University of Alberta

Characterization of pSCL2, A Giant Linear Plasmid in *Streptomyces clavuligerus*

by

Wei Wu ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirement
for the degree of Doctor of Philosophy in Microbiology and Biotechnology

Department of Biological Sciences

Edmonton, Alberta
Fall 2003

Your file Votre référence
ISBN: 0-612-88069-9
Our file Notre référence
ISBN: 0-612-88069-9

University of Alberta

Library Release Form

Name of Author: Wei Wu

Title of Thesis: Characterization of pSCL2, A Giant Linear Plasmid in *Streptomyces clavuligerus*

Degree: Doctor of Philosophy

Year this Degree Granted: 2003

Date submitted to the FGSR: Sep. 29, 2003

Dedicated to all those who have always been there for me...

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Characterization of pSCL2, A Giant Linear Plasmid in *Streptomyces clavuligerus*" submitted by Wei Wu in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Microbiology and Biotechnology.

Dr. Kenneth L. Roy, supervisor

Dr. Susan E. Jensen, co-supervisor

Dr. Laura S. Frost

Dr. Gwen Allison

Dr. George Chaconas, external examiner

Date that the thesis is approved by committee:

September 26, 2003

ABSTRACT

Streptomyces are Gram-positive filamentous soil bacteria, with high GC content in their DNA (70-74%). One of the remarkable features of this genus is that most species have a linear chromosome and many have linear plasmids, as well as circular plasmids. They are thus distinguished from most other bacteria that have circular chromosomes and plasmids. *Streptomyces clavuligerus,* a producer of β-lactam antibiotics, harbours three linear plasmids, pSCL1 (11.7 kb), pSCL2 (120 kb) and pSCL3 (460 kb). In order to increase our understanding of the biology of the linear plasmids in *Streptomyces*, this research has focused on determination of the DNA sequence and characterization of genes of pSCL2.

The giant linear plasmid pSCL2 was isolated using three unconventional methods. In order to determine its DNA sequence, two random fragment libraries of pSCL2 were constructed, which contained small inserts (1 to 3 kb) and larger inserts (3 to 6 kb) respectively. The random "shotgun" strategy was used in the early stages of the sequencing project, while several other strategies, such as primer-walking and nested-deletions, were used in the later stages. Approximately 80% of the nucleotide sequence of pSCL2 has been determined, of which 83 kb was assembled into large contigs. The overall GC content of the assembled sequence is 69.97%. A total of 98 open reading frames (ORFs) were predicted and annotated through sequence analysis and database searching. Genes encoding proteins with putative replication, segregation, regulation and transfer functions were found in pSCL2. Approximately half of the ORFs (49 of them) have no obvious homology to known genes in Genbank. Computational analysis of the central region of pSCL2 revealed the presence of two replication genes, *repC1* and *repC2*. They encode proteins that are highly homologous to the replication proteins of pSLA2-L, the large linear plasmid in *Streptomyces rochei*. Functional analysis indicated that the replication protein RepC1 is essential for replication initiation of pSCL2, whereas RepC2 may be involved in stability or copy number control. The putative replication origin is located approximately 1000 bp upstream of *repC1*.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Dr. Kenneth L. Roy, for providing me with this opportunity to work on this project, and for his continuous advice and encouragement throughout my graduate program. I have learned so much, not only from his rich knowledge and invaluable experience, but also from his enthusiasm to science and life. I would like to thank my co-supervisor, Dr. Susan E. Jensen. She is such an intelligent and understanding supervisor. Her dedication to research is definitely inspiring. I would also like to thank my supervisory committee member, Dr. Laura S. Frost, for her guidance and suggestions towards my research project.

I sincerely appreciate Patricia Murray and Lisa Ostafichuk in the Molecular Biology Service Unit for running numerous sequencing samples for me. My project cannot be accomplished without your cooperation. I also wish to thank all the support staff in the Department of Biological Sciences for being so helpful and keeping everything running smoothly.

Many thanks to my colleagues: Vania Clementino in Dr. Roy's lab, Annie Wong, Cecilia Anders, Liru Wang, Hui Cai, Jingru Li and Kapil Tahlan in Dr. Jensen's lab, and many others in the research group of Microbiology & Biotechnology. Thank you all for sharing your friendship and technical expertise with me, which made my life in these five years much easier and more enjoyable!

I would like to express my deep gratitude to my parents, grandparents, younger brother, and all the people who have being caring me. Thank you for your love, understanding, and faith in me in my life. Especially, I thank my love Weiwu Chen with all my heart for his commitment and unconditional support of me. You are the wind under my wing!

Finally, I acknowledge University of Alberta for providing the FS Chia Scholarship to me, and the Department of Biological Sciences for the Graduate Teaching Award.

TABLE OF CONTENTS

CHAPTER 2 MATERIALS AND MATHODS

CHAPTER 3 RESULTS

CHAPTER 4 DISCUSSION

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

aa	Amino acid
ARS	Autonomously Replicating Sequence
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BLASTN	Nucleotide similarity search using BLAST
BLASTP	Amino acid similarity search using BLAST
bp	Base pair(s)
BSA	Bovine Serum Albumin
CHEF	Contour-clamped Homogeneous Electric Field
CIP	Calf Intestinal Phosphatase
dCTP	2'-deoxycytidine-5'-triphosphate
7-deaza dGTP	2'-deoxy-7-deazaguanosine-5'-triphosphate
dGTP	2'-deoxyguanosine-5'-triphosphate
DMSO	Dimethylsulfoxide
DNA	Deoxyribose Nucleic Acid
dNTP	2'-deoxy(adenosine/thymidine/cytidine/guanosine)-5'-triphosphate
DTT	1,4-Dithiothreitol
dsDNA	Double-stranded DNA
EDTA	Ethylenediamine tetraacetic acid
Exo	Exonuclease
IS	Insertion Sequence
kb	Kilobase
kDa	Kilodalton
LB	Luria-Bertani (medium)
Mb	Megabase
Mw	Molecular weight

nt	Nucleotide
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PEG	Polyethylene glycol
PFGE	Pulse Field Gel Electrophoresis
pI	Isoelectric point
PSI-BLAST	Position Specific Iterative - BLAST
RBS	Ribosome Binding Site (Sequence)
RNase	Ribonuclease
SDS	Sodium Dodecyl Sulfate
SNA	Soft Nutrient Agar
ssDNA	Single-stranded DNA
SSPE	Standard Saline-Phosphate-EDTA
TB	Terrific Broth (medium)
TE	Tris-EDTA buffer
TEA(TAE)	Tris-EDTA Acetate buffer
TEB	Tris-EDTA Borate buffer
TIR	Terminal Inverted Repeat(s)
T_m	Melting temperature
TP	Terminal Protein
TPR	Tetratrico Peptide Repeat
TSB	Tryptic Soy Broth
TSBS	Tryptic Soy Broth-Starch
UV	Ultraviolet
w/v	Weight per volume
YEME	Yeast Extract-Malt Extract

CHAPTER 1 INTRODUCTION

The genus *Streptomyces* is a group of Gram-positive, high GC content, filamentous soil bacteria. It belongs to the family of Streptomycetaceae, which further belongs to actinomycetes, based on morphology and subsequently cell wall chemotype (Anderson & Wellington, 2001). This genus has been extensively studied in the last 50 years because of the complex morphology of its members and their ability to produce a great variety of secondary metabolites, which include at least 60% of known, naturally occurring antibiotics. Many of these antibiotics are widely used in medicine and agriculture. The DNA of *Streptomyces* has a high GC content, as does that of most other actinomycetes. The other remarkable feature of this genus is that most species have a linear chromosome and many have linear plasmids, as well as circular plasmids. They are thus distinguished from most other bacteria that contain circular chromosomes and plasmids. This feature provides a model system for understanding genome structure, genetic behavior, DNA replication, and bacterial evolution of the linear genetic elements.

This chapter first summarizes aspects of current research on *Streptomyces* biology, and then focuses on studies of *Streptomyces* chromosomes and accessory genetic elements. The third part will present a review of bacterial plasmids, including circular and linear plasmids, and their replication mechanisms. Finally, the objectives of this study will be discussed.

1.1 CURRENT RESEARCH

Five aspects of research are introduced in this section, fairly covering the current studies on *Streptomyces* biology. All of these fundamental studies are helpful to understand *Streptomyces* as a complex, prokaryotic genetic system, as well as important to understand the control of secondary metabolite production and the related morphological differentiation.

1.1.1 Antibiotic Biosynthesis and Metabolism Control

It is known that actinomycetes produce about two-thirds of antibiotics that are produced by all microorganisms, and amongst them, nearly 80% are produced by the members of *Streptomyces* (Kieser *et al.*, 2000). Knowledge of the organization of the genes, the intermediates in the pathways of biosynthesis, the regulatory mechanisms, including the switch from primary to secondary metabolism, has been accumulated for a number of antibiotics. Development of versatile cloning vectors, gene transfer systems, and other genetic tools have facilitated genetic manipulation to construct antibiotic-producing strains with improved properties (Baltz, 1998).

In addition to antibiotics, *Streptomyces* also produce a variety of other classes of biologically active secondary metabolites, such as antitumor compounds, anitfungal agents, herbicides, extracellular degradative enzymes and their inhibitors. When it was isolated, *Streptomyces clavuligerus* was notable for the production of two new antibiotics, penicillin N and cephamycin C (Higgens & Kastner, 1971). Now this species is known to produce many more secondary metabolites besides these two antibiotics, including other β-lactams with antimicrobial and antifungal activity, two exocellular proteases and two characterized β-lactamase inhibitors that can contribute to a useful strategy for overcoming the emergence of β-lactam antibiotic resistant organisms.

1.1.2 Morphological Differentiation

Streptomycetes have very complex morphology, which conforms to the large size of their chromosomes, and is also one of the reasons for studying their genetics. Abundant regulators control the complex life cycle, as well as other changes coincident with differentiation, such as onset of secondary metabolism and production of some extracellular proteins. The functions, interactions and expression of regulatory genes, their control of the metamorphosis of aerial hyphae into spore chains, and the significance of the synthesis of sporulation-association proteins and pigments, are topics that have attracted considerable interest.

Under suitable conditions, germ tubes emerge from spores and grow into substrate mycelium. After about 2-3 days, aerial hyphae grow up in a process that involves a large number of *bld* genes. The first *bld* gene, *bldA*, was originally cloned in 1985 (Piret & Chater, 1985), and sequenced in 1987 (Lawlor *et al.*, 1987). More *bld* genes were discovered and studied in the last decade. A model for a regulatory signal cascade in *S. coelicolor* summarized the relationship of these *bld* genes (Kelemen & Buttner, 1998). Intracellular signals and extracellular factors involved in the initiation of aerial mycelium have been studied as well. At least two different pathways, in which SapB and streptofactin are involved, respectively, lead to the formation of aerial hyphae (Kelemen & Buttner, 1998).

The conformation of the aerial hyphae is then further changed. In *S. coelicolor*, the apical compartment of individual aerial hyphae coils to a spiral syncytium containing multiple genomes, and later is subdivided into unigenomic pre-spore compartments. The pre-spores then turn into mature spores after wall thickening and grey pigment deposition. A number of *whi* genes are involved in regulation of the sporulation process. One of the *whi* genes, *whiG*, encodes a sigma factor; and mutation of this gene blocks spore formation (Chater *et al.*, 1989). Similar processes are presumed to occur in other *Streptomyces* spp.

1.1.3 Instability and Evolution

Genetic instability of *Streptomyces* has been known for many years. Initially, it was observed that several genetic markers, such as antibiotic biosynthesis and resistance, pigment and extracellular protein production, and aerial hyphae and spore formation, were lost in various recombinations (Volff & Altenbuchner, 1998). Such genetic instability has been shown to result from frequent rearrangements of the linear chromosomal DNA, including deletions, amplifications, inversions, integrations, chromosomal arm replacements, and circularizations. Many of these phenomena are related to the presence of linear plasmids and their interaction, like recombination, with linear chromosome in their host cells. Recent work has shown that deletions and circularizations often occur in the terminal sequences of the linear chromosomes, which sometimes contain transposable elements and other horizontally transferred elements; the replicative transposition caused by these transposable elements may be a major source of the chromosomal instability (Chen, 1996; Chen *et al.*, 2002). The underlying mechanism of the instability is not clear yet, but Rec proteins have been suggested to be involved in the chromosome amplification and control of genetic instability (Volff & Altenbuchner, 1998).

The structural instability is evidently related to evolution of the *Streptomyces* chromosomes. Volff and Altenbuchner (2000) proposed that linear *Streptomyces* chromosomes could result from recombination between linear plasmids and ancestral circular chromosomes, and this might have happened relatively recently.

1.1.4 Replication

Three main hypotheses have been put forward to explain the replication of linear chromosomes and plasmids (details will be discussed in Section 1.3.2). A mechanism consisting of bi-directional initiation from a centrally located *oriC* and fold-back replication of terminal sequences is most likely to describe the replication mode of *Streptomyces* linear chromosomes and plasmids.

The replication origins of many *Streptomyces* chromosomes have been studied (Lin *et al.*, 1993; Calcutt 1994; Lezhava *et al.*, 1995; Leblond *et al.*, 1996; Fischer *et al.*, 1998; Lu *et al.*, 2002). They share conserved gene organization. Certain short sequences are conserved in the origin region and have been demonstrated to be essential for initiation of replication (Zakrzewska-Czerwinska *et al.*, 2000). Few replicons in Streptomycete linear plasmids have been thoroughly investigated. Studies on the replication origins of two small linear plasmids, pSCL1 from *S. clavuligerus* and pSLA2-S from *S. rochei*, showed a similar pattern of direct repeats in their origin regions (Chang *et al.*, 1996). Other functional replication origins of linear *Streptomyces* plasmids that have been studied include pSLA2-L in *S. rochei* (Hiratsu *et al.*, 2000), SCP1 in *S. coelicolor* (Redenbach *et al.*, 1999), and SLP2 in *S. lividans* (Huang *et al.*, 2003). These origin regions comprise a series of direct and/or inverted repeat sequences, low GC content sequence, and genes encoding replication proteins, but do not show obvious homologies in their sequences.

Terminal sequence patching is another important topic in the study of *Streptomyces* linear replicons. In this process, both the terminal proteins and the end sequences in the terminal inverted repeats (TIR) play essential roles, which were found to be relatively conserved between chromosomes and plasmids, and within different species of *Streptomyces* (Huang *et al.*, 1998; Bao & Cohen, 2000; 2003). Further studies on the replication origins and telomeres of the *Streptomyces* linear replicons will present a clearer picture for the understanding of their replication mechanism.

1.1.5 Chromosomes and Accessory Genetic Elements

Physical and genetic maps of the chromosomes and plasmids of some *Streptomyces* spp. have been constructed by pulse field gel electrophoresis and analysis of ordered libraries. Some of these *Streptomyces* chromosomes and plasmids have been completely sequenced (Table 1-1, Table 1-2). Plasmids, bacteriophages, and other mobile elements have been characterized, some of which are becoming useful tools for genetic manipulation. Each genetic element in *Streptomyces* will be discussed individually in the next section.

The availability of the recently completed genome sequence of *S. coelicolor* (Bentley *et al.*, 2002), the model organism for streptomycetes and other filamentous actinomycetes, is likely to have an important impact on all aspects of the research introduced above. This linear chromosome is among the largest (8.7 Mb) and highest in GC content (72%) of bacterial chromosomes. A wealth of genes expanded from the ancestral core allows the organism to have a more complex life cycle, a larger secondary metabolite repertoire, and to adapt to a wider range of environmental conditions. This is in contrast to the *Borrelia burgdorferi* linear chromosome, which is among the smallest (910 kb) and lowest in GC content (28.6%) of bacterial chromosomes (Fraser *et al.*, 1997). The small size of the *B. burgdorferi* genome reflects its simple parasitic lifestyle. The abundant genes in *Streptomyces* also provide a greater potential for its broader applications in the fields of medicine, agriculture and other industries.

Table 1-1. Summary of some *Streptomyces* linear chromosomes.

Streptomyces spp.	Size (Mb)	Antibiotics the chromosome produces	References
S. coelicolor A3(2) *	8.67	Actinorhodin Prodiginines Calcium dependent antibiotic	Bentley *et al.*, 2002
S. clavuligerus	8	Cephamycin C Penicillin N	Higgens & Kastner, 1971
S. lividans 66	8	Actinorhodin	Leblond *et al.*, 1993
S. rimosus R6-501	8	Oxytetracycline	Pandza *et al.*, 1997
S. avermitilis ATCC31267 *	9.03	Avermectin	Omura *et al.*, 2001; Ikeda *et al.*, 2003
S. ambofaciens	8	Spiramycin	Leblond *et al.*, 1991
S. griseus IFO03237	7.8	Streptomycin	Lezhava *et al.*, 1995

* Completely sequenced

Table 1-2. Summary of some linear and circular plasmids in *Streptomyces*.

Plasmids	Size (kb)	Host	Antibiotics the plasmid produces	References
pSCL1*	11.7	*S. clavuligerus*	-	
pSCL2	120	NRRL 3585	-	Netolitzky *et al.*, 1995;
pSCL3	460		-	this study
SCP1*	350	*S. coelicolor* A3(2)	Methylenomycin	Bentley *et al.*, 2002
SCP2* #	31		-	Haug *et al.*, 2003
pSLA2-S	17	*S. rochei*	-	Kinashi *et al.*, 1994
pSLA2-M	100		-	Kinashi *et al.*, 1994
pSLA2-L*	210.6		Lankacidin; Lankamycin	Kinashi *et al.*, 1994 Mochizuki *et al.*, 2003
SAP1 *	94.3	*S. avermitilis*	-	Omura *et al.*, 2001;
pSA2	250		-	Evans *et al.*, 1994
pPZG101	387	*S. rimosus*	-	Pandza *et al.*, 1998
pPZG103	1000		Oxytetracycline	"
pSB1	640	*S. bambergiensis*	-	Zotchev *et al.*, 1992
pBL1	43	S712	-	"
pSV1 #	175	*S. violaceoruber*	Methylenomycin	Spatz *et al.*, 2002
pSV2 *	97	SANK95570	-	"
pSJL1	11.7	*S. jumonjinesis*	-	Netolitzky *et al.*, 1995
pSJL2	17.5	NRRL 5741	-	"
pSJL3	220		-	"
pSJL4	280		-	"
pSGL1	120	*S. griseus* NRRL 3851	-	Netolitzky *et al.*, 1995
SLP2 *	50	*S. lividans* 1326	-	Huang *et al.*, 2003

* The plasmids that have been completely sequenced.

# Circular plasmids; the remaining unlabelled plasmids are all linear.

1.2 *STREPTOMYCES* CHROMOSOMES AND ACCESSORY GENETIC ELEMENTS

One of the most remarkable features of *Streptomyces* is that most species have a linear chromosome and many have linear plasmids, as well as, in some cases, circular plasmids. They are thus distinguished from most other bacteria that contain circular chromosomes and plasmids.

As members of the high GC content group of the Gram-positive bacteria, *Streptomyces* species have among the highest proportions of these residues in their chromosomal DNA, often quoted as 70-74% (Kieser *et al.*, 2000), and sometimes with an even wider range. This high GC content results in a strong, biased distribution of GC composition at the third position of codons within a protein-coding region (Bibb *et al.*, 1984). Normally, for a given protein-coding region, the GC percentage at the three positions in codons is about 70% at the first position, 50% at the second position and over 90% at the third. This characteristic distribution has been used as a criterion to identify structural genes isolated from *Streptomyces*. Because of the high GC content, it has often been observed that restriction endonucleases recognizing sites low in GC cut *Streptomyces* DNA at lower frequencies than those enzymes recognizing sites high in GC content. Moreover, it is has been suggested that the high GC content may reduce the possibility of UV damage.

1.2.1 Chromosomes

Streptomyces was originally presumed, like other typical prokaryotes, to contain a circular chromosome, until about 10 years ago. Two free ends found in the *Streptomyces lividans* chromosome by restriction mapping (Lin *et al.*, 1993) led to the discovery that most streptomycetes chromosomes have a linear structure (Lezhava *et al.*, 1995). The accurate size of the *Streptomyces* genome was hard to determine until pulsed-field gel electrophoresis (PFGE) was used to analyze large restriction fragments of the chromosome (Kieser *et al.*, 1992). It revealed that *Streptomyces* have among the largest chromosomes found in bacteria, in the range of 6 to 9 Mb. The large genome reflects the complex life cycle and large secondary metabolite repertoire of these free-living soil microbes. Table 1-1 lists some *Streptomyces* species that have been well investigated. Recently the complete genome sequences of two *Streptomyces* species, *S. coelicolor* and *S. avermitilis*, have been determined (Bentley *et al.*, 2002, Ikeda *et al.*, 2003). About 5,300 genes, 68% of their total protein-coding genes, of the two Streptomycetes were found to be closely related. The remaining genes are apparently unique to each genome. This is not surprising since almost all the genes encoding enzymes for secondary metabolites are different between these two species (Hopwood, 2003).

The linear chromosomes have a centrally located replication origin (Fischer *et al.*, 1998; Musialowski *et al.*, 1994), and terminal inverted repeats (TIR) with terminal proteins covalently bound to the 5'-ends of the DNA (Lin *et al.*, 1993; Leblond *et al.*, 1996). The sizes of the chromosome TIRs varies from 24 kb in *S. griseus* 2247 (Lezhava *et al.*, 1995) to 550 kb in *S. rimosus* R6-501 (Pandza *et al.*, 1997). A rare naturally occurring exception is that a 174-bp TIR was found in the chromosome of *S. avermitilis* ATCC31267 (Omura *et al.*, 2001). This could be possibly caused by the exchange between the ends of the chromosome and the ends of a linear plasmid that contains a short TIR (Hopwood, 2003). The sequences of TIRs are quite diverse in different species, except for about the last 160 bp towards the ends (Huang *et al.*, 1998).

Chromosome linearity was thought to be causally related to another hallmark of *Streptomyces* chromosomes – instability, which refers to their tendency to suffer large deletions, amplifications and circularizations. However, more recent studies showed that *in vivo* artificially circularized chromosomes turned out to be even more unstable than their linear parents and further deletions were observed (Lin & Chen, 1997). However, after a further deletion, a few of the circular survivors became relatively stable (Lin & Chen, 1997). Such observations imply that the terminal sequences that contained transposable elements and were deleted in the circularization process probably play a more important role in chromosomal instability than its linearity. Terminal chromosomal deletions may involve both ends and may extend up to nearly a quarter of the chromosome. Leblond *et al.* (1991) even suggested that an atypical strain of *S.*

ambofaciens, which had a genome size of only 6.5 Mb, was perhaps a spontaneous deletion variant. Genetic instability occurs spontaneously at very high frequencies (more than 0.1%) in the chromosomes of most *Streptomyces* species, and even more frequently after certain treatments (Kieser *et al.*, 2000).

1.2.2 Plasmids

Many prokaryotes and eukaryotes contain plasmids, extrachromosomal genetic elements that are able to replicate independently. Circular and linear plasmids are two general types of bacterial plasmids. Many *Streptomyces* strains contain linear plasmids, as well as circular plasmids in some cases. Linear plasmids are also found in a wide range of other actinomycetes regardless of the topology of their host chromosomes. The first *Streptomyces* plasmid identified and physically isolated was the circular plasmid SCP2 from *S. coelicolor* (Schrempf *et al.*, 1975), and the first reported linear plasmids were pSLA1 and pSLA2, isolated from *S. rochei* 7434-AN4 (Hayakawa *et al.*, 1979).

SCP2*, a large circular plasmid spontaneously derived from SCP2, has been fully sequenced (Haug *et al.*, 2003; Table 1-2). This 31.3 kb, low copy number plasmid was originally identified in *S. coelicolor* A3(2) as a fertility factor. It probably replicates by the "theta" bi-directional mode. Smaller circular plasmids tend to have higher copy numbers. One extensively studied small circular plasmid, pIJ101 (8.9 kb) from *S. lividans*, has 300 copies per genome. It has been developed as an important vector for use in *Streptomyces* genetics. The complete nucleotide sequence of pIJ101 has been reported (Kendall & Cohen, 1988). This small plasmid replicates autonomously by a rolling-circle method via a circular single-stranded replication intermediate (Zaman *et al.*, 1993). Some other *Streptomyces* circular plasmids, like SLP1 (17 kb) of *S. coelicolor* and pSAM2 (11 kb) of *S. ambofaciens*, are integrating plasmids. They normally reside in the host chromosome, but transfer independently by conjugation to other strains lacking the plasmids, where they can turn into autonomously replicating circular plasmids.

The known linear plasmids in *Streptomyces* are between 10 kb and 1 Mb in size (Table 1-2). The large ones that are over 50 kb in size are often referred to as giant linear plasmids. Like the chromosome, they have TIRs and terminal proteins covalently bound at the 5' ends. The sizes of TIRs in linear plasmids vary from 44 bp in SLP2 (50 kb) of *S. lividans* (Chen *et al.*, 1993) to 95 kb in pPZG101 (387 kb) of *S. rimosus* (Gravius *et al.*, 1994). They replicate in a similar way to the mode used by the linear chromosomes in their hosts. Several *Streptomyces* linear plasmids have been completely sequenced, including pSCL1 (11.7 kb) from *S. clavuligerus* (Wu & Roy, 1993), SCP1 (350 kb) from *S. coelicolor* (Bentley *et al.*, 2002), pSV2 (97 kb) from *S. violaceoruber* (direct summission to NCBI Genbank), SLP2 (50 kb) from *S. lividans* (Huang *et al.*, 2003), and pSLA2-L (210.6 kb) from *S. rochei* (Mochizuki *et al.*, 2003).

Most plasmids in *Streptomyces* do not contain antibiotic biosynthetic genes or antibiotic resistance determinants except for the following cases: SCP1 in *S. coelicolor* (Kirby *et al.*, 1975; Kinashi *et al.*, 1987) and circular plasmid pSV1 in *S. violaceoruber* (Aguilar & Hopwood, 1982) encoding for methylenomycin; pPZG103 in *S. rimosus* encoding for oxytetracycline (Pandza *et al.*, 1998); and pSLA2-L in *S. rochei* encoding for lankacidin and lankamycin (Suwa *et al.*, 2000).

Most *Streptomyces* plasmids detected so far are conjugative and associated with "pock" formation, which refers to macroscopically visible, circular areas of retarded growth that develop around colonies growing from individual plasmid-carrying spores seeded in a lawn of plasmid-free spores (Kieser *et al.*, 2000). Conjugation provides bacteria with the capacity for horizontal gene transfer, in most cases within one species or two closely related organisms. One example is that the linear plasmid SLP2 can be transferred among *S. lividans*, *S. coelicolor* and *S. parvulus*. Plasmid-mediated conjugation gives actinomycete soil bacteria the potential to share useful genetic information, such as heavy-metal resistance, antibiotic production and antibiotic resistance. The *Streptomyces* circular and linear plasmids use a different mechanism during transfer. Possoz *et al.* (2001) proposed that circular plasmid pSAM2 of *S. ambofaciens* is probably transferred in double-stranded form. However, it is suggested that transfer of the *Streptomyces* linear plasmids starts from the end of the molecules in a single-stranded form (Chen, 1996). In *Streptomyces* the only identified conjugation related genes were reported in pBL1 of *S. bambergiensis* (Zotchev & Schrempf, 1994).

1.2.3 Transposable Elements

A number of transposable elements, including insertion sequences and transposons, have been discovered in streptomycetes (Kieser *et al.*, 2000; Kieser & Hopwood, 1991), but none of them naturally contain selectable markers (Kieser *et al.*, 2000).

Insertion sequences (IS) are the simplest transposable elements, usually encoding one protein required for transposition. IS are able to transpose into new locations in DNA present in a cell, either within or between replicons (Iida *et al.*, 1983). IS*110* in *S. coelicolor* was the first IS-like element in *Streptomyces* to be physically isolated (Chater *et al.*, 1985). This 1.6 kb segment is able to transpose from the *S. coelicolor* chromosome into ϕC31 phage derivatives, with an *attP* deletion. The terminal sequence of IS*110* showed short imperfect inverted repeats. IS*110* belongs to the IS*110* family (Mahillon & Chandler, 1998), which also includes a previously discovered insertion sequence IS*116* in *S. clavuligerus* (Leskiw *et al.*, 1990). The members in the IS*110* family have diverse structures and host strains, which range from species of *Streptomyces*, *Mycobacterium, Rhodococcus*, *Lactobacillus*, *Moraxella*, to vancomycin-resistant enterococci.

Most IS-like elements exist in a linear form. One of the exceptions is IS*117*, a 2.6 kb "mini-circle" found in *S. coelicolor* (Lydiate *et al.*, 1986; Henderson *et al.*, 1990). When introduced into *S. lividans* 66, a species closely related to *S. coelicolor* but lacking IS*117* and other mini-circle sequences, IS*117* integrates into the host chromosome at specific sites with a circular transposition intermediate (Henderson *et al.*, 1990). Derivatives of this element containing resistance markers have been used as integrating vectors (Kieser *et al.*, 2000).

The first reported *Streptomyces* transposon, Tn*4556*, was found in a neomycin-producing strain of *Streptomyces fradiae* in 1987 (Chung, 1987). It was discovered when it jumped onto the plasmid prophage SF1 in that strain. Sequence data revealed that Tn*4556* is 6.8 kb in length, with 68% GC content. It has 38 bp terminal inverted repeats, which are 70% identical to the Tn*3* transposon in *E. coli*. In addition, it encodes a 97 kDa transposase and a 16 kDa resolvase, which are 61% and 45% identical to the transposase and resolvase encoded on Tn*3*. Transposable elements that randomly and stably insert into DNA *in vivo* can be very useful for generalized mutagenesis and gene tagging. Tn*4556* and IS*493* of *S. lividans* have been developed as tools for random mutagenesis of *Streptomyces*.

1.2.4 Bacteriophages

Many bacteriophages, both temperate and virulent, have been found in *Streptomyces* species, but they rarely infect other genera of actinomycetes (Kieser *et al.*, 2000). The best studied *Streptomyces* phage is ϕC31, which is from *S. coelicolor* and has a broad host range (Chater, 1986). The genome sequence of ϕC31 has been fully determined (Smith *et al.*, 1999), and has a length of 41.4 kb with a 63% GC content. The DNA has cohesive ends that result from short, single-stranded complementary termini, just like λ phage in *E. coli*. Vectors made from ϕC31 and other temperate phages, such as R4, VP5, SH10, have been important tools for cloning and manipulating *Streptomyces* DNA.

1.3 PLASMIDS AND THEIR REPLICATION

In general, plasmids are considered to be a group of double-stranded DNA elements that do not contain genes essential for cell survival. Two main conformations have been identified: circular form and linear form. On the basis of their DNA structures, different plasmids employ different replication modes. In this section, various types of circular and linear plasmids, as well as the replication mechanisms that are used by each type of plasmid will be introduced. The discussion here will emphasize bacterial plasmids, but may refer to eukaryotic plasmids and other extrachromosomal elements when it is helpful and applicable.

1.3.1 Circular Plasmids

There are three primary replication mechanisms for circular plasmids: "theta" bi-directional replication, strand displacement, and rolling circle. Some linear elements that circularize prior to replication also use these methods (Lewin, 1987).

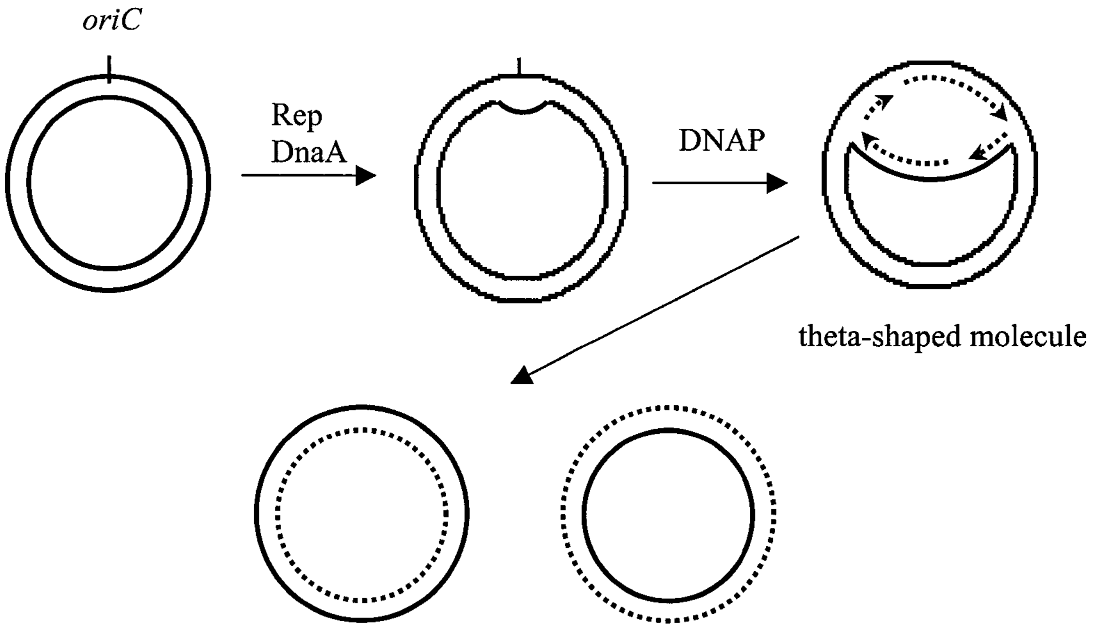
Replication by the theta-type mechanism is extensively used among plasmids from Gram-negative and Gram-positive bacteria, as well as most bacterial circular chromosomes (Madigan *et al.*, 2003). This mechanism was named because the replication intermediates appear as bubbles in the early stage, resulting in theta-shaped molecules. Replication can start from one or several origins, and it can be uni- or bi-directional. This mode of replication initiates from opening of the strands at the specific site (*oriC*), which is catalyzed by Rep and DnaA proteins. Unwinding enzymes and topoisomerases further catalyze unwinding of the double strands. Meanwhile, RNA primers are synthesized either by RNA polymerase or by bacterial/plasmid primases. DNA molecules are then synthesized at the growing end of the RNA primers. DNA synthesis of the two strands is coupled, occurring continuously on the leading strands and discontinuously on the lagging strands. The leading strands are primed only once at the beginning, however, each short stretch of DNA in the lagging strands requires an individual RNA primer. DNA polymersae III is required for elongation of DNA molecules, while DNA polymerase I participates in the early synthesis of the leading strands in some cases (ColE1 and pAMβ1 plasmids) and replaces RNA primers with DNA in the lagging strands (del Solar *et al.*, 1998). This mechanism of replication is illustrated in Fig. 1-1A.

The best-known examples of plasmids replicating by the strand displacement mechanism are the IncQ family plasmids (as typified by RSF1010). Fig. 1-1B outlines a model for initiation of RSF1010 replication, proposed by Scherzinger *et al.* (1991). Replication occurs from two symmetrical and adjacent origins (*ssiA* and *ssiB*), which are positioned one on each strand, and independently used. Replication starts when these origins are exposed as single-stranded regions. The melting of the DNA strand depends on two plasmid replication proteins, RepC and RepA, and is facilitated by an AT-rich region that precedes the *ssiA* and *ssiB*. RepA is a DNA helicase, and RepC recognizes the direct repeats in the origin that is adjacent to the AT-rich region. Priming of DNA

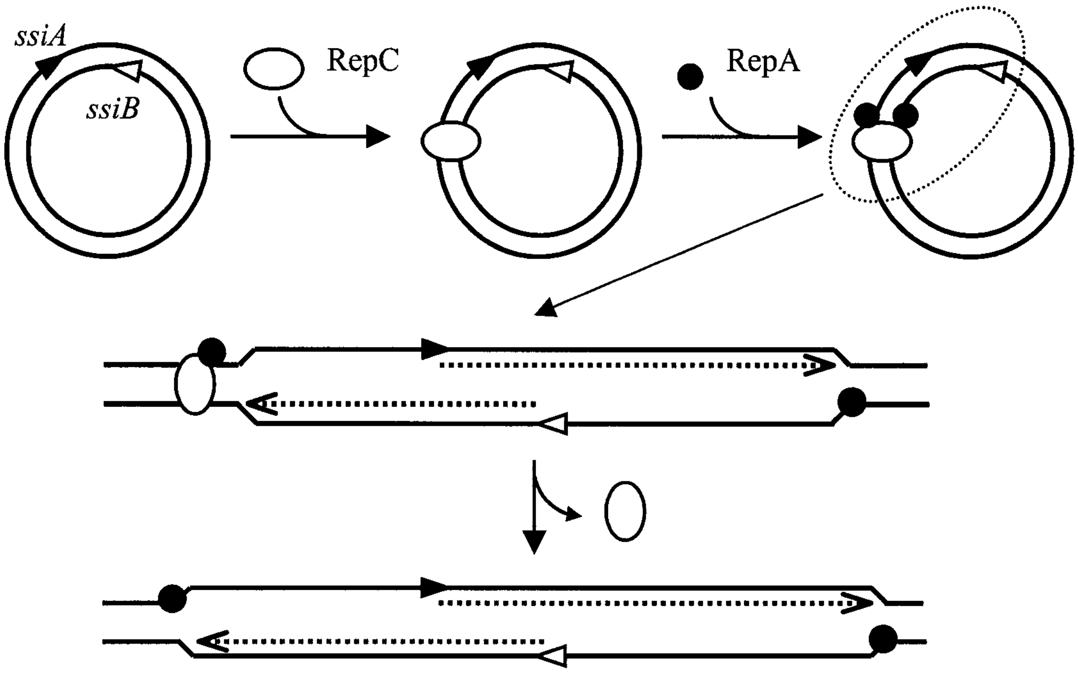
Fig. 1-1. Diagrams depicting circular plasmid replication.

(A) Theta type replication; (B) strand displacement replication; (C) rolling circle replication. Solid and dotted lines indicate parental and daughter strands, respectively. The identities of other components are indicated by appropriate labels. DNAP, DNA polymerase; RNAP, RNA polymerase. Other details are given in the text.

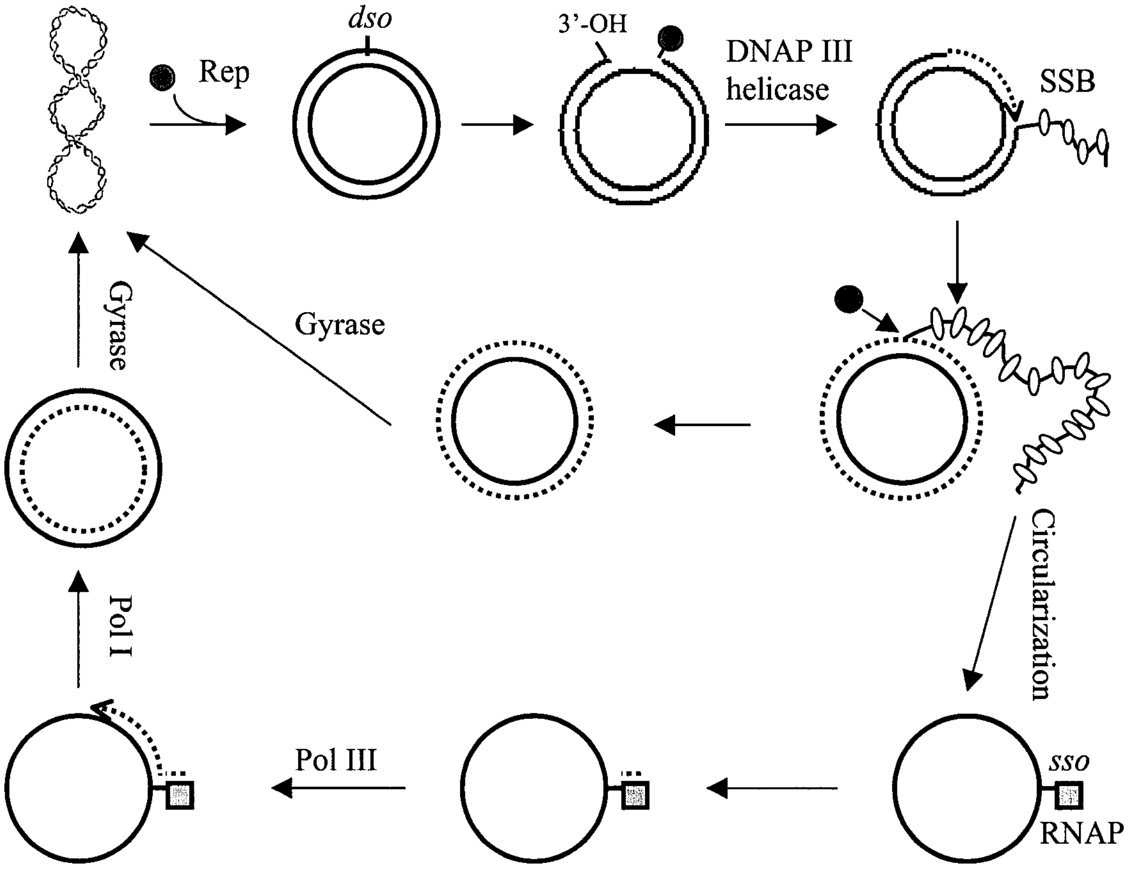
A. Theta-type bi-directional DNA replication



B. Strand displacement DNA replication (based on the diagram in del Solar *et al.*, 1998)



C. Rolling circle DNA replication (based on the diagram in del Solar *et al.*, 1998)



synthesis at these origins is catalyzed by the plasmid-specific primase (RepB, not shown in the figure). Synthesis of each one of the strands occurs continuously and results in the displacement of the complementary strands.

Replication by the rolling circle mechanism is widespread among multicopy plasmids from the *Archaea* and *Bacteria*. The current model includes several stages (Fig. 1-1C). Replication is initiated by the plasmid-encoded Rep protein, which unwinds the supercoiled DNA and introduces a nick at the double-stranded origin (*dso*) on the plus strand. The nick leaves a 3'-OH end that is used as a primer for leading-strand synthesis. Elongation from the 3'-OH end, accompanied by the displacement of the parental plus strand, involves host replication proteins, at least including DNA polymerase III, helicase and single-stranded DNA binding proteins (SSB). Leading-strand synthesis is terminated by various specific strand transfer reactions, also mediated by the Rep proteins. After completion of leading-strand synthesis, the Rep protein is inactivated. The end products are a dsDNA molecule constituted of the parental minus strand and the newly synthesized plus strand, and a ssDNA intermediate corresponding to the parental plus strand. The ssDNA intermediate is turned to dsDNA by lagging-strand synthesis. It is initiated by the host RNA polymerase (RNAP) at the single-strand origin (*sso*) that is physically distant from the *dso*. RNAP produces a short primer RNA, from where lagging-strand DNA is synthesized by host DNA polymerase. At last, the host DNA gyrase converts the close-circled, relaxed DNA into supercoiled form (del Solar *et al.*, 1998).

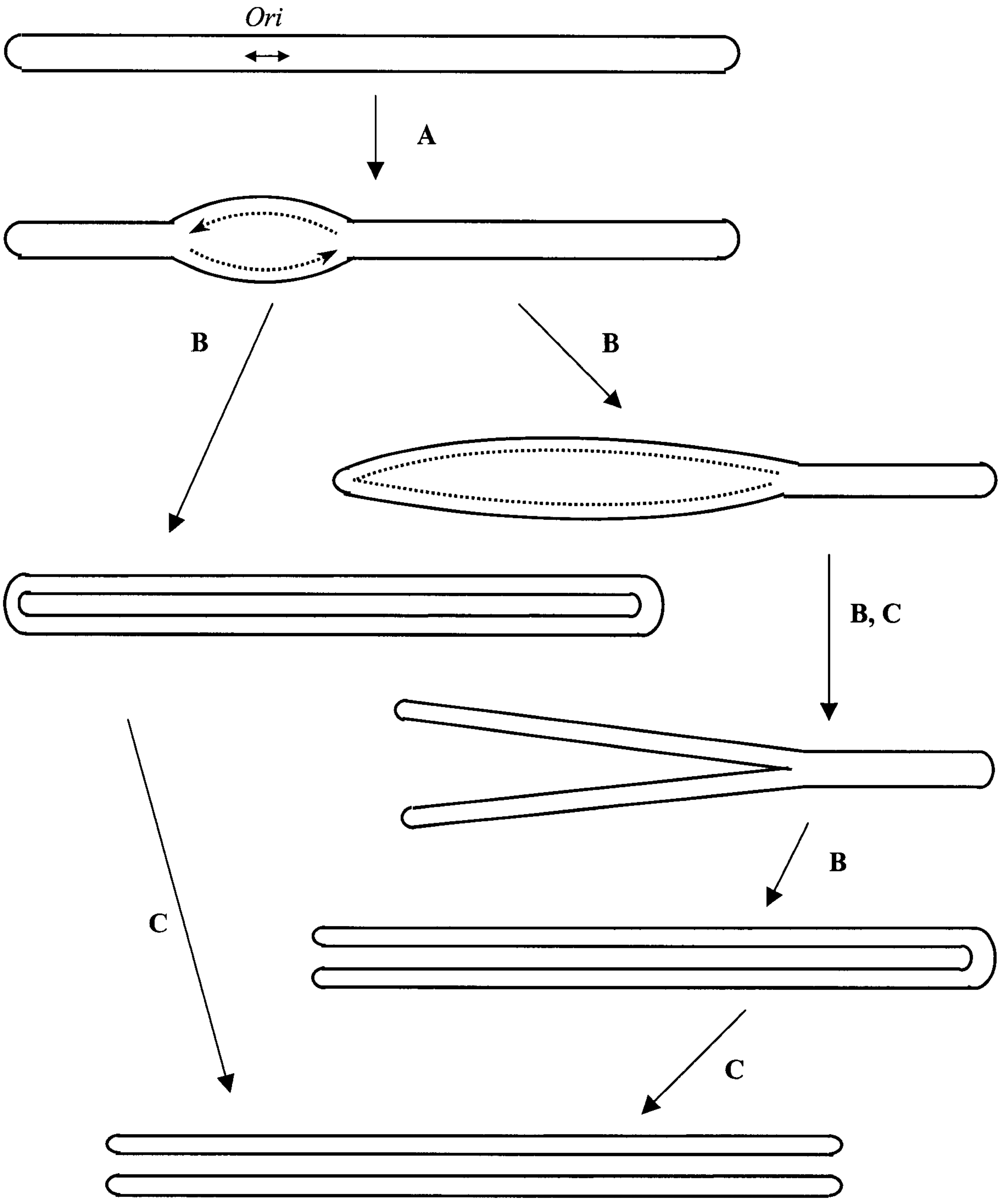
1.3.2 Linear Plasmids

Two structural types of linear plasmids have been characterized, the so-called hairpin plasmids with covalently closed ends and invertron plasmids with a protein covalently attached to their 5' terminus at each end.

Hairpin plasmids are common in human-pathogenic *Borrelia* spirochetes, such as *Borrelia hermsii* that causes relapsing fever, and *Borrelia burgdorferi* that causes Lyme disease. Another example of a hairpin plasmid is *E. coli* linear phage-plasmid N15. Both the *Borrelia* plasmids and N15 serve as models for studying the mechanism of replication of these linear replicons. The hairpin plasmids have short terminal inverted repeats (TIR) and covalently closed hairpin loops at both termini (Hinnebusch & Barbour 1991). Recent investigations on *Borrelia* plasmids suggested that initiation of bi-directional DNA replication occurs internally, proceeding to the ends, where a circular intermediate is formed. This was supported by the physical mapping of the replication origin and DNA sequence analysis of the linear plasmids for AT and GC skew (Picardeau *et al.*, 2000b). The circular intermediate is subsequently broken by a "DNA breakage and reunion reaction", in which linear DNA molecules with hairpin ends are regenerated (Chaconas, *et al.*, 2001). The authors showed that a synthetic sequence with a structure of a linear plasmid telomere is able to function as a substrate for telomere resolution *in vivo*; a circular replicon with such a synthetic sequence can also be converted into a linear molecule (Chaconas, *et al.*, 2001). A similar replication mode was also discovered in N15 (Ravin, 2003). When the replication origin is not located in the centre of the linear molecule, one end may finish duplication earlier, thus a Y-shaped structure is formed. After the duplication of the other end is completed, two linear molecules are produced (Ravin, 2003). A schematic representation of this replication mechanism is shown in Fig. 1-2. In addition, phage-plasmid ϕK02 in *Klebsiella oxytoca* and pY54 in *Yersinia enterocolitica* seem to use a similar mechanism for their replication (Ravin, 2003). Additional hairpin type linear replicons are found in eukaryotes, such as poxviruses, African swine fever virus, some *Chlorella* viruses, and certain yeast mitochondrial plasmids, however, these use a different replication mechanism (Ravin 2003; not shown in diagram). Specifically, poxvirus replication is initiated in the telomere, resulting in the formation of head-to-head and tail-to-tail concatemers through strand-displacement; the replicated telomeres in the concatemers are subsequently resolved by an as yet unknown enzyme to generate monomeric molecules with covalently closed ends (Traktman, 1996).

DNA plasmids with proteins covalently linked to their 5' ends comprise the largest group of extrachromosomal linear elements. All elements in this group have terminal inverted repeats (TIR), varying in size from 44 bp (SLP2 in *S. lividans*; Chen *et al.*, 1993) to 95 kb (pPZG101 in *S. rimosus*; Gravius *et al.*, 1994), so for this reason these

Fig. 1-2. Diagrams depicting replication of *Borrelia* linear plasmids with hairpin ends. Solid and dotted lines indicate parental and daughter strands, respectively. The identities of other components are indicated by appropriate labels. Other details are given in the text.



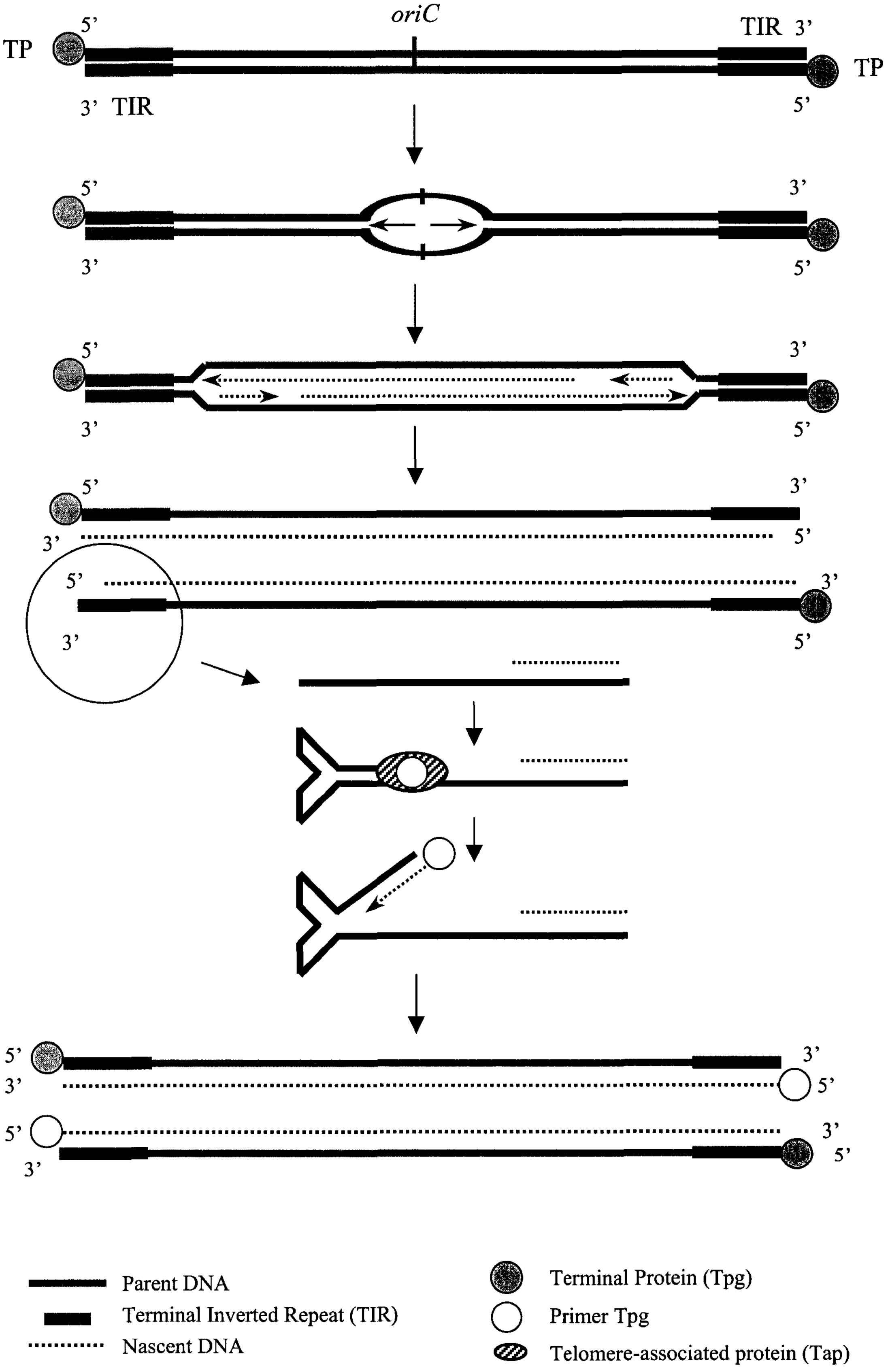
A. Replication initiation
B. Replication
C. Telomere breakage and reunion

genetic elements are also known as invertrons (Sakaguchi, 1990). Most of the bacterial linear plasmids of this kind are found in Actinomycetes, although there have been a few reported from non-Actinomycete bacteria. In addition, they are located in organelles, usually mitochondria, of plants and fungi (Meinhardt *et al.*, 1990; Griffiths 1995), and in the cytoplasm of yeast (Stark *et al.*, 1990). The invertron-like structure is also found in some bacteriophages, adenoviruses, and transposons (Sakaguchi, 1990; Hinnebusch & Tilly, 1993).

Current studies on the replication of extrachromosomal linear elements have led to the proposal of two primary mechanisms. One based on the linear plasmids and chromosomes in *Streptomyces* (Fig. 1-3) suggests that the linear replicons replicate bidirectionally from a centrally located origin towards the telomeres (Chang & Cohen, 1994; Chang *et al.*, 1996; Fischer *et al.*, 1998). The first evidence supporting this proposal came from studies showing that *Streptomyces* linear plasmids do not require telomeres for replication and that derivatives of the linear plasmid pSCL1 in *S. clavuligerus* can replicate as circular forms when the telomeres are removed and the resulting ends are ligated (Shiffman & Cohen, 1992). The location of the replication origin was later confirmed by the two-dimensional agarose gel electrophoresis patterns of the restriction fragments of replicating pSLA2-S in *S. rochei* (Chang & Cohen, 1994). The replication intermediates of the bi-directional mechanism are linear duplex molecules that have short recessed 5' ends (approximately 280 nucleotides in pSLA2-S in *S. rochei*), and leave single-strand gaps at the 3' ends (Chang & Cohen, 1994). Several possible mechanisms, including a fold-back model, a recombination model and a terminal hairpin model, for the filling-in of recessed 5' ends have been proposed by Chen (1996). Later results from Qin and Cohen's study (1998) supported the fold-back mechanism, in which the 3' leading-strand overhang folds back, due to the presence of multiple terminal palindromes, resulting in a DNA duplex, thereby providing a recognition site for the terminal protein that serves as a primer to complete DNA synthesis (patching) at the terminus and produce full-length double-stranded DNA molecules. The detailed process and the proteins involved in the terminus patching will be discussed in Section 1.3.4.

Fig. 1-3. Diagrams depicting replication of *Streptomyces* linear plasmids.

Solid and dotted lines indicate parental and daughter strands, respectively. The identities of other components are indicated by appropriate labels. Other details are given in the text.



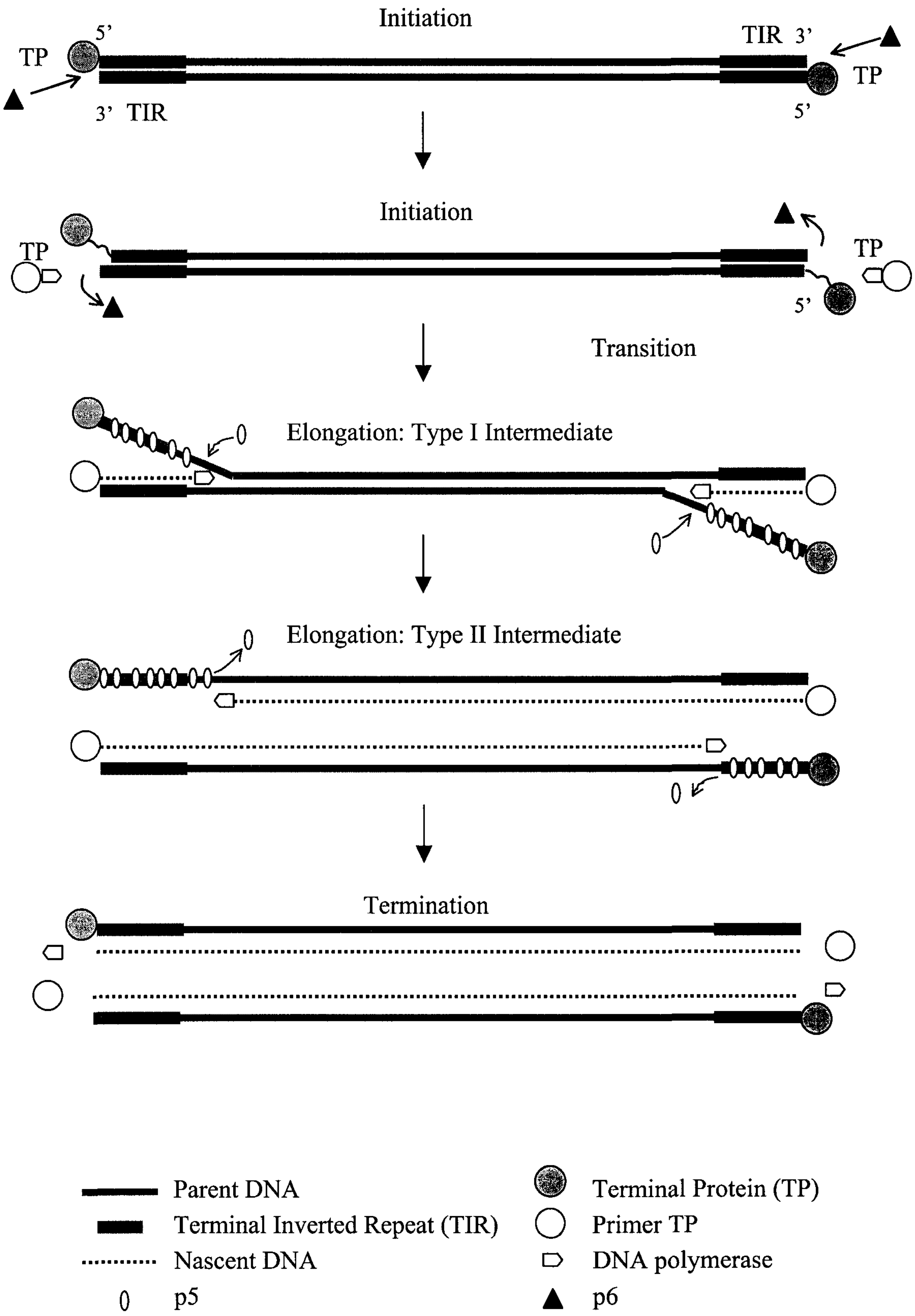
oriC
TP
TIR
5'
3'
3' TIR
TIR 3'
TP
5'
Parent DNA
Terminal Inverted Repeat (TIR)
Nascent DNA
Terminal Protein (Tpg)
Primer Tpg
Telomere-associated protein (Tap)

Bacillus subtilis phage ϕ29 DNA is another extensively studied model for elucidation of the mechanism of linear DNA replication in prokaryotic cells (Salas, 1991; Meijer *et al.*, 2001). The genome of ϕ29 is a 19.3 kb, linear, double-stranded DNA. It contains a terminal protein (TP) covalently linked at each 5' end, which constitute the origins of replication. The ϕ29 protein-primed replication (Fig. 1-4) initiates by recognition of the origin by a heterodimeric protein complex formed by the primer TP and the phage-encoded DNA polymerase. The virus-encoded protein p6 binds to double-strand DNA to help open the DNA ends (Serrano *et al.*, 1994), facilitating the addition of the first nucleotide to the primer TP, which is catalyzed by the ϕ29 DNA polymerase (Blanco & Salas, 1984). After a transition step, these two proteins dissociate, and the DNA polymerase continues processive elongation. The nascent DNA strand displaces the other parent strand, which is bound by the protein p5 (single strand binding protein (SSB)) at the same time. The replication starts at both DNA ends so that the coupled strand displacement generates so-called type I replication intermediates consisting of full-length double-stranded ϕ29 DNA molecules with one or more ssDNA branches of varying lengths. When the two DNA polymerase replication complexes meet, a type I intermediate physically separates into two type II replication intermediates, each containing a full-length parental DNA strand and an incomplete daughter strand. DNA synthesis continues until replication of the nascent DNA strand is completed. Such a mechanism is also employed by the *E. coli* phage PRD1 (Caldentey *et al.*, 1993), the *Streptococcus pneumoniae* phage Cp-1 (Martín *et al.*, 1996), and eukaryotic adenoviruses (van der Vliet, 1995; de Jong & van der Vliet, 1999).

Besides the above two main mechanisms, eukaryotic linear plasmids evolve a strategy that uses an RNA template and the riboenzyme telomerase for telomeric DNA synthesis (McEachern *et al.*, 2000; Blackburn, 2001). Other linear replicons also apply "turnaround" replication to accomplish lagging-strand DNA synthesis and consequently contain a single-strand loop located between inverted repeat sequences at the ends of duplex DNA (Kornberg & Baker 1992) (not shown in diagram).

Fig. 1-4. Diagrams depicting protein-primed replication of linear DNA.

Solid and dotted lines indicate parental and daughter strands, respectively. The identities of other components are indicated by appropriate labels. Other details are given in the text.


Initiation
TP
5'
3' TIR
TIR 3'
TP
5'
Initiation
TP
TP
5'
Transition
Elongation: Type I Intermediate
Elongation: Type II Intermediate
Termination
Parent DNA
Terminal Inverted Repeat (TIR)
Nascent DNA
p5
Terminal Protein (TP)
Primer TP
DNA polymerase
p6

1.3.3 Partitioning

Genes whose products direct the proper partitioning of daughter molecules following replication are not only located on chromosomes for normal growth and development of cells, but have also been discovered on low-copy number plasmids to prevent plasmid loss during segregation. In general, *par* loci are organized as gene cassettes, and almost all known plasmid-encoded *par* loci consist of three components: a centromere-like site and two genes. The upstream gene encodes an ATPase that is essential to the DNA segregation process, and the downstream gene encodes a DNA-binding protein that binds to the centromere-like region (Gerdes *et al.*, 2000). Two types of partitioning ATPases are known, one that contains the Walker-type ATPase motif (Koonin, 1993) and the other that belongs to the actin/hsp70 superfamily of ATPases (Bork *et al.*, 1992). Based on the type of ATPase encoded, all known *par* loci are divided into two families, type I and type II. The type I loci can be further divided into Ia and Ib subgroups, depending on the genetic organization and gene sizes within the loci (Gerdes *et al.*, 2000).

So far, all of the partitioning genes found in *Streptomyces* belong to type I, and most of them are type Ia loci. In type Ia *par* loci, the two genes, *parA/sopA* and *parB/sopB*, encode ParA/SopA and ParB/SopB proteins, respectively. The ATPases (ParA/SopA) are in a size range of about 250-420 amino acids and contain DNA-binding domains at their N-termini (Hayes *et al.*, 1994). The ParB/SopB proteins are about 250-340 aa in size, and bind as dimers to the centromere-like site *parS/sopC* that is located downstream of and adjacent to *parB/sopB* (Davis & Austin, 1988; Mori *et al.*, 1989; Surtees & Funnell, 1999). The ParA/SopA proteins do not bind directly to *parS*, but control transcription of the *par* operons via binding to operator sequences in the promoter regions (Mori *et al.*, 1989; Davis *et al.*, 1992; Hayes *et al.*, 1994).

In type Ib *par* loci, ParA and ParB proteins are significantly smaller (190-300 aa and 50-130 aa, respectively) than the partitioning proteins in type Ia, and the *parS* site coincides with the promoter region upstream of the *parA* and *parB* genes (Gerdes *et al.*, 2000). Many type Ib loci have multiple direct repeats both upstream and downstream of the *par* genes. In contrast to the type Ia proteins, no DNA-binding domains are found in the type Ib ParA ATPases. ParB proteins of type Ib loci autoregulate the *par* operons via binding to the centromere-like regions that contain the *par* promoters (Gerdes *et al.*, 2000).

The type II *par* loci contain two genes *parM* and *parR*, encoding ParM (actin-like ATPase) and ParR respectively. A centromere-like region named *parC* is located upstream of *parM* and *parR* genes and overlaps with the *par* promoter (Dam & Gerdes, 1994; Gerdes & Molin, 1986). ParR proteins repress the *par* promoter by binding cooperatively to the iterons in the *parC* site, whereas ParM is not involved in regulation (Gerdes *et al.*, 2000).

The characteristics of type Ia, Ib and type II partitioning loci are summarized in Table 1-3.

1.3.4 Telomeres

It has been known for more than 20 years that *Streptomyces* linear plasmids have terminal proteins covalently attached to their 5' ends, which protect these ends from degradation by exonucleases (Hirochika & Sakaguchi, 1982). Later it was found that the *Streptomyces* linear chromosomes also have terminal proteins with similar features. However, until recently the characteristics of the terminal proteins were unknown. Bao and Cohen (2001) revealed that the terminal protein (Tpg) is encoded by the *tpg* gene, and further discovered that a telomere-associated protein (Tap) that had previously been identified as a putative transcriptional regulator has an essential role in 5' end replication (Bao & Cohen, 2003). Tpg and Tap proteins are unique in linear replicons in *Streptomyces* spp and have not been found in any other organisms.

Streptomyces linear plasmids and chromosomes were originally thought to replicate by the protein-primed full-length strand-displacement mechanism that is used by adenovirus and phage ϕ29 (Salas, 1991; Salas *et al.*, 1995). It is now known that they replicate bidirectionally from an internal site near the centre of the molecule (Shiffman & Cohen, 1992; Chang & Cohen, 1994; Musialowski *et al.*, 1994), during which replicative

Table 1-3. Summary of the characteristics of type Ia, Ib and type II *par* loci.

	ATPase (essential)	Centromere -binding protein	Centromere-like site
Type Ia	ParA/SopA Walker-type ATPase Large (250-420 aa) Control operon transcription by binding to operator sequences in the promoter	ParB/SopB Large (250-340 aa) Bind as dimmers to *parS/sopC*	*parS/sopC* Downstream of *parAB/sopAB* genes
Type Ib	ParA Walker-type ATPase, probably lack DNA-binding domain Smaller (190-300 aa) Do not autoregulate its own synthesis	ParB Bind to *parS* Very small (50-130 aa) Autoregulate the operon via binding to the promoter-*parS* overlapped region	*parS* Upstream of *parAB* genes Overlapped with *par* promoter
Type II	ParM Actin-like ATPase Not involved in the regulation	ParR Repress the *par* promoter by binding to the iterons in the *parC* site	*parC* Upstream of *parMR* genes Overlapped with *par* promoter

intermediates containing a 3' overhang of leading-strand DNA are produced (Chang & Cohen, 1994; Fig. 1-3). In order to generate blunt-ended DNA molecules telomere patching is required, in which Tpg and Tap proteins are involved.

Both the terminal protein Tpg and the telemere-associated protein (Tap) are essential for *Streptomyces* plasmids and chromosomes to replicate as linear molecules. Cells that survive after disruption of *tpg* or *tap* genes show telomere deletion and circularization of the plasmids and chromosomes. It is suggested that the Tap protein interacts with both the Tpg protein and a specific sequence at the 3' overhangs, and positions Tpg to the terminus (Bao & Cohen, 2003; Fig. 1-3). Tpg then probably binds to the recognition site formed by the 3' overhangs, and leads to prime synthesis of the lagging-strand 5' terminus (Qin & Cohen, 1998; 2000; Fig. 1-3).

The DNA sequences of the telomeres of *Streptomyces* linear replicons contain inverted repeats (palindromes), and are quite different from the telomeres of eukaryotic chromosomes. Huang *et al.* (1998) compared the terminal segments of several *Streptomyces* linear chromosomes and linear plasmids, including chromosomes of *S. lividans*, *S. coelicolor*, *S. parvulus* and *S. lipmanii*, and plasmids SLP2 from *S. lividans*, pSPA1 from *S. parvulus*, pSCL1, pSCL2 from *S. clavuligerus* and pSLA2-S from *S. rochei*. The last 166-168 bp segment of their terminal sequences shows a high degree of conservation but most sequences beyond do not share significant homology. Moreover, the termini of pSCL1 and pSLA2-S are quite similar to each other, but have lower similarities to the termini of other investigated plasmids and chromosomes. The homologous regions of pSCL1 and pSLA2-S contain seven palindromes. Deletion of these widely spaced palindromic DNA sequences in the telomeres directs *Streptomyces* linear replicons to replicate in a circular form.

1.4 BACKGROUND AND OBJECTIVES OF THIS STUDY

Streptomyces clavuligerus, isolated in 1971 from a South American soil sample (Higgens and Kastner, 1971), is one of the *Streptomyces* species that has been extensively studied in the last thirty years. It can produce a variety of β-lactam antibiotics and β-lactamase inhibitors. *S. clavuligerus* contains three linear plasmids, pSCL1 (11.7 kb;

Keen *et al.*, 1988; Wu & Roy, 1993), pSCL2 (120 kb) and pSCL3 (430 kb; Netolitzky *et al.*, 1995). The small linear plasmid pSCL1 was first discovered in *S. clavuligerus* in 1988 (Keen *et al.*, 1988). Since the other two linear plasmids had not been found at that time, pSCL1 was thought to be the only plasmid in *S. clavuligerus* and named as pSCL. Five years later, Wu and Roy (1993) determined the complete nucleotide sequence of pSCL1. It was the first linear plasmid to be fully sequenced in *Streptomyces*. The medium and large linear plasmids, pSCL2 and pSCL3, were later discovered by Netolitzky *et al.* (1995), and their sizes, as estimated by pulse field electrophoresis gel (CHEF), were about 120 kb and 430 kb.

The major objectives of this study include three aspects: determination of the nucleotide sequence of pSCL2, the second linear plasmid in *S. clavuligerus*; identification of the possible protein-coding regions on this plasmid; and analysis of its replication activity. There are a number of reasons to conduct this research.

Firstly, *S. clavuligerus* is of medical and industrial importance due to its ability to produce β-lactam antibiotics, including cephamycin C and cephalosporin, as well as β-lactamases and β-lactamase inhibitors, including clavulanic acid. β-lactamase inhibitors can be used for overcoming the emergence of β-lactam antibiotic resistant organisms. It has been reported that some *Streptomyces* giant linear plasmids are involved in biosynthesis of secondary metabolites, like SCP1, pSLA2-L and pPZG103. The sequence study of pSCL1 has shown that this small linear plasmid does not carry any genes for antibiotic production (Wu & Roy, 1993). It was thought to be worthwhile to investigate whether pSCL2 is involved in the biosynthesis of any secondary metabolites in *S. clavuligerus*.

Secondly, compared with what is known about the replication origins of *Streptomyces* chromosomes (Lin *et al.*, 1993; Calcutt 1994; Lezhava *et al.*, 1995; Fischer *et al.*, 1998; Lu *et al.*, 2002), the replication features of most Streptomycete linear plasmids are poorly understood. Certain common features, such as short conserved sequences and similar gene organization, have been found in *Streptomyces* chromosome origin regions and shown to be essential for initiation of replication (Zakrzewska-Czerwinska *et al.*, 2000). The small linear plasmid pSLA2-S has been the model replicon for studying the replication mechanism of linear DNA molecules in *Streptomyces* (Chang *et al.*, 1996). Only two large linear plasmids, SCP1 (Redenbach *et al.*, 1999) and pSCLA2-L (Hiratsu *et al.*, 2000), have been thoroughly studied for their replication origins. However, no common features were found between them. Sequence information for pSCL2 is essential to identify the replication origin(s) of this giant linear plasmid, which will no doubt enrich our current knowledge on replication of *Streptomyces* linear plasmids.

Lastly, DNA sequence data is uniquely useful to reveal important information that cannot readily be obtained by other approaches, including the GC content and its variation within the complete sequence, the overall gene organizations, the presence of repeated elements or insertion elements, the identification of DNA regions acquired by horizontal gene transfer and other new operons, etc.

Two sequencing strategies have frequently been used (Frangeul, *et al.*, 1999). The first one is direct shotgun sequencing, which is also called the random fragmentation method. In this method DNA is sheared randomly into small pieces using mechanical forces such as sonication, non-specific enzyme digestion, and nebulization. The random shotgun approach has become a popular strategy for sequencing because it does not require preliminary data (such as a map) before the sequencing phase. It has been applied in several genome sequencing projects, such as *Helicobacter pylori* (Tomb *et al.*, 1997), *Streptococcus pyogenes* (Ferretti *et al.*, 2001) and *S. avermitilis* (Omura *et al.*, 2001). The second strategy is the ordered-clone approach that uses a large-insert library to construct a map of overlapping clones covering the whole genome; then selected clones are sequenced to obtain the whole-genome sequence. The advantage of the ordered-clone approach is that it makes the closure phase easier since fewer gaps may be left. This method was used in the sequencing projects for the *S. coelicolor* genome and linear plasmid pSLA2-L, in which ordered cosmid libraries were constructed before sequencing (Redenbach *et al.*, 1996; Mochizuki *et al.*, 2003). Considering the advantages of each of these two strategies, both were worth trying for this project.

The method for sequence analysis was a matter of program and software selection. The DNAstar (DNASTAR) software package was used for sequence assembly and general sequence analysis since it was available in our laboratory. Sequencher software (Gene Codes) is efficient, particularly for sequence assembly, so that it can be used to confirm the assembly results obtained from DNAstar. GeneTool (BioTools) was also available for general sequence analysis. As discussed earlier, *Streptomyces* DNA has an unusually high GC content, which results in a characteristically strong and biased GC distribution in the protein-coding regions. FramePlot is a program designed to take advantage of this feature of the *Streptomyces* genome for gene prediction. BLAST (Basic Local Alignment Search Tool; http://www.ncbi.nlm.nih.gov/BLAST/; Altschul *et al.*, 1990) and FASTA (http://www.ebi.ac.uk/fasta33/; Pearson & Lipman, 1988) are the two most popular homology search engines to explore all of the available sequence databases for both DNA and protein sequence query. A more sensitive tool for protein comparisons, PSI-BLAST (Position specific iterative BLAST; Altschul *et al.*, 1997), combines statistically significant alignments produced by BLAST into a position-specific score matrix. This iterative searching strategy effectively increases sensitivity for weak but biologically relevant sequence similarities so as to be widely used for annotation of hypothetical proteins with low conservation. With the boom of bioinformatics in recent years, more and more tools have been developed for sequence analysis, structure prediction and gene annotation, which are too extensive to be summarized and introduced here. The choice of the tools applied in this study was made on a case-to-case basis through the whole project.

CHAPTER 2 MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 Bacterial Strains

Streptomyces clavuligerus NRRL 3585 was obtained from S.E. Jensen, but originated from the North Regional Research Laboratories (NRRL), Peoria, Illinois. *Streptomyces lividans* TK24 was obtained from T. Kieser, John Innes Institute, Norwich, England.

TOP10 One Shot® *E. coli* (F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L(Str^R) *end*A1 *nup*G) was used as the host strain for electroporation, and was purchased from Invitrogen, Inc. The JM109 chemically competent *E. coli* cells (*end*A1, *rec*A1, *gyr*A96, *thi*, *hsd*R17 (r_k^-, m_k^+), *rel*A1, *sup*E44, Δ(*lac-pro*AB), [F', *tra*D36, *pro*AB, *lac*I^{q}ZΔM15]) were purchased from Promega Corporation. The nonmethylating *E. coli* stain ER1447 (*dam*$^-$, *dcm*$^-$) was provided by S.E. Jensen, Dept. of Biological Sciences, University of Alberta.

2.1.2 Plasmids

The cloning vectors, pCR®4Blunt-TOPO® and pCR®-BluntII-TOPO®, were purchased from Invitrogen, Inc. The pGEM7Zf(+) and pGEM11Zf(-) vector were obtained from Promega. pHJL400 was provided by S.E. Jensen.

2.1.3 Culture Conditions and Growth Media

All *E. coli* strains were maintained in glycerol stocks stored at -70°C. For plasmid preparation, *E. coli* cultures were grown overnight in LB (Luria-Bertani) or TB (Terrific Broth) medium (Sambrook *et al.*, 1989) containing kanamycin (50 μg/ml) or ampicillin

(100 μg/ml) at 37°C in roller tubes. For generation and regeneration of electro-competent cells, *E. coli* cells were grown and prepared as described in Section 2.4.1.

All *Streptomyces* strains were maintained as spore stocks in glycerol at -70°C. For sporulation, *S. clavuligerus* was grown on TOA (Tomato/Oatmeal Agar; 20 g/L tomato paste, 20 g/L finely ground oatmeal, 25 g/L agar, pH 6.5) for 2 weeks, after which the spores were harvested as described in Kieser *et al* (2000). For isolation of genomic and plasmid DNA from *Streptomyces* spp, spores were inoculated in TSBS (Tryptic Soy Broth (DIFCO) with 1% soluble starch) liquid medium to grow mycelium. All cultures were incubated at 28°C in flasks containing wire springs, with shaking at 280 rpm for 2 to 3 days until a dense culture was obtained. For protoplast preparation, *Streptomyces* cells were grown as described in Section 2.4.2.

2.1.4 Enzymes, Chemicals and Supplies

Restriction endonucleases and DNA modifying enzymes were purchased from Roche and New England Biolabs (NEB). All enzymes were used as recommended by the manufacturers. *Taq* and *Pfu* polymerases for PCR were obtained from M.A. Pichard, Dept. Biological Sciences, University of Alberta.

The antibiotics ampicillin, kanamycin and neomycin were obtained from Sigma, St. Louis, MO. Chemicals and reagents used for standard molecular biology research techniques were purchased from either Sigma, St. Louis, MO; ICN, Aurora, Ohio or BDH, Toronto, Ontario. Growth media were purchased from DIFCO Laboratories, Detroit, Michigan. [α-^{32}P] labeled dCTP was obtained from Amersham, Arlington Heights, IL. [γ-^{32}P] labeled ATP was obtained from ICN Biochemicals Inc.

Random sequence oligonucleotides were purchased from Boehringer Mannheim. Oligonucleotides used as primers for DNA sequencing and PCR were synthesized in the MBSU (Molecular Biology Services Unit) in the Department of Biological Sciences, University of Alberta, on the Applied Biosystems 392/391 DNA Synthesizer using standard phosphoramidite chemistry.

Nylon membranes (Hybond-N) used for Southern transfers were purchased from Amersham. Kodak X-ray film was purchased from the Eastman Kodak Co.

2.2 DNA PREPARATION

2.2.1 Preparation of Genomic DNA from *Streptomyces*

2.2.1.1 *In situ* Cell Lysis

In order to prepare intact genomic DNA, including chromosomal DNA and giant linear plasmids, from *S. clavuligerus*, *in situ* cell lysis and DNA isolation were accomplished by embedding the cells in agarose gel to prevent DNA shearing (Fahnert *et al.*, 2000). Mycelium was harvested by centrifugation and washed with 4 M guanidine hydrochloride and 0.2 M EDTA in HE buffer (10 mM Hepes-NaOH, 1m M EDTA, pH 8.0) and centrifuged. The pellet was then resuspended in HES buffer (25 mM Hepes-NaOH, 25 mM EDTA, 0.3 M sucrose, pH 8.0) and mixed with low melt point agarose (made in HES) at 45°C to reach a final concentration of agarose of 0.5 to 0.8%. The suspension was poured into a Bio-Rad CHEF sample plug mould and allowed to solidify at 4°C. The plugs were removed from the moulds and incubated with gentle shaking in HES buffer containing 5 mg/ml lysozyme for 3 hours at 37°C. Then they were treated with proteinase K (0.5 mg/ml) in NDS buffer (10 mM glycine, 0.5 M EDTA, 1% SDS, pH 9.5) overnight at 37°C. Finally the blocks were treated with Pefabloc (0.5 mg/ml; Boehringer Mannheim) in HE buffer for 1 hour at room temperature. It was necessary to wash the plugs with the buffer to be used in the next step between each two steps. The plugs were washed with HE buffer at 4°C for 4 times, for 30 minutes each time, with frequent shaking, then stored in HE buffer with 100 to 500 mM EDTA at 4°C.

2.2.1.2 Quick Method for Total DNA Isolation

It normally took two days to prepare DNA samples when using the previously described *in situ* cell lysis method. However, a quick approach for total DNA isolation was used if very high molecular weight intact DNA molecules were not required, such as when preparing DNA for PCR templates. A small amount of mycelium from 5 ml 40-hour culture was harvested and washed twice in 0.3 M sucrose. The pelleted mycelium was suspended in 700 µl TES buffer (25 mM Tris-HCl, 25 mM EDTA, 0.3 M sucrose, pH 8.0) containing 2 to 4 mg/ml lysozyme and 50 µg/ml RNase, and incubated for 30 to 45 minutes at 37°C. This was followed by mixing with 70 µl 10% SDS and extraction using an equal volume of a phenol and chloroform mixture four times. DNA was then precipitated from the aqueous phase by adding three volumes of ethanol, and spooled with a sealed Pasteur pipette. The precipitated DNA was washed with 70% ethanol and dissolved in TE buffer.

2.2.2 Preparation of Linear Plasmid DNA from *Streptomyces*

2.2.2.1 Isolation of DNA from Agarose Gel Embedded Cells

The large linear plasmids *of S. clavuligerus*, pSCL2 and pSCL3, could be separated from chromosomal DNA by Contour-clamped Homogeneous Electric Field (CHEF; Chu *et al.*, 1986; Section 2.3.1) electrophoresis of DNA embedded in standard or low melting point agarose gel and isolated from the agarose gel by electroelution (Section 2.3.3.2).

2.2.2.2 Isolation of Linear Plasmid DNA by Sucrose Gradient Centrifugation

An approach developed by Drs. D.J. Netolitzky and K.L. Roy for isolation and separation of giant linear plasmids was also used to prepare high molecular weight DNA and purify pSCL2 (Netolitzky *et al.*, 1995). In this method the mycelium was gently digested with lysozyme to form protoplasts, followed by gentle treatment with proteinase K and SDS.

The harvested *Streptomyces* mycelium was washed and resuspended in TE buffer with 15% sucrose. Lysozyme and additional EDTA were added to reach a final lysozyme concentration of 1 mg/ml and 100 mM EDTA. After incubation at room temperature for one hour, proteinase K was added (to a final concentration of 0.4 mg/ml), mixed by very gentle inversion, followed by the addition of 20% SDS (to a final concentration of 1.5%), gentle inversion, and incubation at 55°C for 3 hours. A NaCl solution was then added (to a final concentration 1 M) and mixed by gentle inversion to clear the lysate. The mixture was chilled on ice then centrifuged at 40,000 rpm at 4°C for 45 minutes in a Ti 50 rotor (Beckman, Beckman Instruments Inc.). The top half of the supernatant was removed, using a wide-mouth micropipette tip. Portions of the supernatant were then loaded onto a 20-50% sucrose gradient for isolation of pSCL2.

The sucrose gradient was made by mixing 6.3 ml 20% sucrose and 6.2 ml 50% sucrose solution (in 1 M NaCl and 2 × TE) in a SW40 tube (total volume as 12.5 ml), using a two-chamber gradient maker. Samples of 0.6 ml total DNA were loaded on the top of the gradient, and centrifuged in a SW 40 rotor at 20,000 rpm for 65 hours to separate molecules larger than 50 kb. Gradients were then fractionated by repeatedly pipetting 0.7 ml fractions from the top. Each fraction was mixed with an equal volume of 2-propanol by gentle inversion and the DNA was allowed to precipitate at -20°C overnight.

2.2.3 Circular Plasmid DNA Preparation from *E. coli* and *Streptomyces*

Plasmids replicated in *E. coli* were isolated and purified according to established methods described by Sambrook *et al.* (1989).

The circular plasmids of *Streptomyces* were isolated using a mini-lysate procedure, which is modified from the above method for circular plasmids in *E. coli*. The modification includes the addition of 2 mg/ml lysozyme in the solution I, followed by incubation of the mixture at 37°C for 30 minutes. The isolated DNA was purified by RNase treatment (50 μg/ml RNase, 50 minutes) and phenol/chloroform extraction.

2.3 DNA MANIPULATION

2.3.1 Conventional and Pulse Field Gel Electrophoresis

DNA fragments in the range of 0.5 to 30 kb were separated, and their sizes determined, using conventional vertical electrophoresis at various concentrations (0.4 to

1.5%) of agarose gel in 1 × TAE buffer (Sambrook *et al.*, 1989). Typical electrophoresis conditions ranged from 1 to 5 V/cm, for 3 to 18 hours.

Higher molecular weight DNA fragments, chromosome and large linear plasmids were separated, and their sizes determined, using the CHEF technique, a form of pulsed field gel (PFG) electrophoresis, which is able to separate DNA fragments up to several megabases in size. CHEF gels were prepared using 1% to 1.5% agarose (either standard or low melting point agarose) in 0.5 × TBE buffer (Sambrook *et al.*, 1989) with 12 mM thiourea, and run in 0.5 × TBE including 60 μM thiourea at 12-15°C in the Bio-Rad CHEF DR II apparatus (Bio-Rad Laboratories). Thiourea was used to avoid Tris-dependent, site-specific cleavage of DNA (Fahnert *et al.*, 2000). Typical pulse settings ranged from 3 to 40 seconds ramp for 30-300 kb DNA fragments, to 40 to 120 seconds ramp for 100 to 800 kb DNA fragments, at 170 V. The molecular weight markers used included Yeast Chromosome PFG Marker (225 to 1900 kb), Lambda Ladder PFG Marker (50 to 1000 kb) and Low Range PFG Marker (0.1 to 200 kb).

All DNA bands on conventional and CHEF gels were stained with ethidium bromide and visualized by illumination with ultraviolet light.

2.3.2 Restriction Endonuclease Digestion of DNA

2.3.2.1 Digestion of DNA in Solution

Restriction enzyme digestion of DNA in aqueous solution was conducted following the conventional protocols (Sambrook *et al.*, 1989), using the manufacturer's recommended conditions and buffers supplied with the enzyme.

2.3.2.2 Digestion of DNA Embedded in Agarose Blocks

Digestion of DNA molecules embedded in agarose gels requires longer time for digestion to allow time for the restriction enzymes to diffuse into the gel. In order to have complete digestion, larger amount of enzymes were also necessary. The gel plugs containing DNA to be digested were washed three times on ice, 30 minutes each time, with the specific digestion buffer in a total volume of 1 ml. The volume of the plug was omitted in the third wash. The DNA was then digested in the smallest possible volume of buffer that was just able to cover the plug. Additional amount of BSA was added into the reaction, to a final concentration of 500 μg/ml, to stabilize the enzymes. Normally 50 to 100 units of enzymes were used to digest DNA in plugs. Incubation was for 3 to 4 hours, followed by addition of an additional 10 to 20 units of enzymes and continuing incubation overnight.

2.3.2.3 Double and Triple Digestion

When the working temperatures and digestion buffers of the enzymes to be used were compatible, all of the enzymes were added in the same reaction simultaneously. The volume of buffer was increased to at least 10 times of the volume of all enzymes added. Otherwise, for digestion in solution, digestion with one enzyme was followed by precipitation of the DNA with ethanol and centrifugation. The pelleted DNA fragments were then redissolved and digested with the second enzyme. For DNA embedded in agarose gel, plugs were washes three times in the specific buffer for the second enzyme, in the same manner as the washes before the first digestion (Section 2.3.2.2), and then the second digest was begun by the addition of the enzyme. The salt concentration of these specific digestion buffers is an important factor in deciding the order of digestions. The enzyme requiring the lowest salt concentration in its specific buffer should be used for the first digestion.

2.3.3 Recovery of DNA from Agarose Gels

2.3.3.1 Digestion of Agarose using Agarase

After separating chromosome and plasmid DNA using CHEF electrophoresis, the gel slices containing pSCL2 and pSCL3, respectively, were cut out and washed twice in at least 2 volumes of 1 × β-agarase I buffer at 4°C, for 30 minutes each. The gel slices were then melted at 65 to 70°C for 10 minutes, immediately followed by cooling to 40°C. The molten agarose, containing pSCL2 or pSCL3, was incubated at 40°C for 2 hours with β-agarase (about 1 unit for each 200 μl 1% agarose). After digestion, DNA was extracted with an equal volume of phenol and precipitated in ethanol at -20°C, overnight.

2.3.3.2 Electroelution of DNA into Dialysis Bags

The dialysis bags (Spectrum Laboratories, Inc.) were treated in advanced by soaking in 2% (w/v) $NaHCO_3$ and 1 mM EDTA solution for 10 minutes at 50 to 55°C, followed by washing several times with distilled water and 1 mM EDTA at 50 to 55°C. The treated membranes were stored in 1 mM EDTA at 4°C and rinsed several times with distilled water, followed by 1 × TEA buffer (50 mM Tris-HOAc, 5 mM EDTA, pH 7.8) at room temperature before use. The bands containing the linear plasmids were excised from CHEF gels, washed with 10 ml of 1 × TEA buffer at 4°C, and the DNA was extracted by electroelution in 1 × TEA, based on the method described by Sambrook *et al.* (1989). After passing an electric current (usually 4-5 V/cm) through the gel slice for 2 to 3 hours, a short reverse of current was carried out for 1 to 2 minutes to disassociate DNA from the membrane before the electrophoresis was finished. The DNA solution removed from the dialysis bags was then concentrated with 1-butanol to the desired volume and the DNA was precipitated with ethanol at -20°C.

2.3.3.3 Glass Milk Purification of DNA

A QIAEX II Gel Extraction Kit (Qiagen, Inc.) was used to extract low molecular weight DNA (fragments of up to 10 kb) from 0.3 to 2% standard or low-melting point agarose gels in TAE or TBE buffer. The procedures followed the manufacturer's instructions.

2.3.4 Determination of the DNA Concentrations

A Unicam UV3 UV/Vis spectrometer (Analytical Technology Inc.) was used to determine DNA concentration. Solution samples were scanned from the wavelengths of

230 nm to 320 nm. The oligonucleotide concentration was calculated using the following formula:

Concentration (μmole/ml) = Peak Absorbance reading (at around 260nm) × 33 × dilution / Mw

A TK100 Mini-Fluorometer (Hoefer Scientific Instruments, San Francisco) was used to determine the concentration of plasmid DNAs. A standard dye solution (working dye solution A), which contains 0.1 μg/ml fluorescent dye Hoechst 33258 solution in 1 × TNE (0.1 M NaCl, 10 mM Tris-HCl, 1 mM EDTA, pH 7.4), was added to samples of DNA to achieve a final concentration between 10 ng/ml and 500 ng/ml. 100 ng/μl calf thymus DNA was used as a standard to calibrate the fluorometer.

2.4 TRANSFORMATION

2.4.1 Preparation of *E. coli* Electro-competent Cells and Transformation

E. coli electro-competent cells were made using the following method. An overnight culture (0.5 ml) was inoculated in LB broth (50 ml). The culture was incubated, with shaking, at 37°C to the mid-log phase, which is considered to have an OD_{600} of 0.35 to 0.5. At this point the cells were chilled completely on ice, harvested by centrifugation for 10 minutes at 3000 × g at 4°C and washed twice thoroughly with ice-cold sterile 10% glycerol to minimize the salt content, using first an equal volume (50 ml), then a one-half volume (25 ml) of 10% glycerol. The final cell pellet was resuspended in 1/50 of the culture volume (1 ml) of ice-cold sterile 10% glycerol, and aliquots of 50 μl were transferred into microfuge tubes, frozen in a dry ice/ethanol bath and stored in a -70°C freezer.

Electroporation was carried out in a 0.2 cm gap electroporation cuvette (Bio-Rad, Fisher Scientific), with a Gene PulserII Electroporation System (Bio-Rad), at 2.5 kV, with a resistance of 200 Ohms and capacitance of 25μFd. *E. coli* cells were regenerated in SOC medium (Sambrook *et al.*, 1989) at 37 °C for 1 hour after transformation before plating on appropriate selective media.

2.4.2 Preparation of *Streptomyces* Protoplasts and Transformation

The preparation, transformation and regeneration of *S. lividans* protoplast were conducted as described by Kieser *et al.* (2000) with certain modifications. A 100 µl sample of protoplasts was thawed and washed once with 500 µl P buffer (protoplast buffer; Kieser *et al.*, 2000) and centrifuged for 2 minutes at 6500 rpm in a microcentrifuge. The pelleted protoplast was resuspended in the last drop of buffer remaining after removal of the supernatant. The DNA to be transformed (up to 20 µl) was added followed by immediately adding 100 µl of T buffer (P buffer containing 25% PEG-1000) and mixing by pipetting up and down twice. P buffer (1 ml) was then added and mixed. The transformed protoplasts in P buffer were centrifuged at 6500 rpm in a microcentrifuge for 2 minutes, resuspended in a suitable volume of P buffer, and gently plated on predried R5 plates. When the surface of the agar developed a faint haze due to growth of the regenerating protoplasts after incubation at 28 to 30°C overnight, the plates were overlaid with soft nutrient agar (SNA) containing 600 µg/ml of neomycin for an effective neomycin concentration of 60 µg/ml. Colonies might be seen after incubation at 28 to 30°C for 2 to 3 days.

2.5 HYBRIDIZATION

2.5.1 Southern Transfers

DNA from conventional gels was transferred to Hybond-N nylon membranes (Amersham) using the Southern technique (Southern, 1975) with certain modifications. The procedure includes depurination for 10 minute in 0.25 M HCl, a 30-minute treatment in denaturing solution (0.5 M NaOH, 1.5 M NaCl), a 30-minute treatment in neutralizing solution (172.6 ml/L HOAc, 2.4 M NaOH, pH 7.4), followed by blotting transfer in the transfer buffer (1 M NH_4OAc, 0.02 M NaOH) for 12-18 hours. The membrane was washed in 2 × SSPE (Sambrook, *et al.*, 1989) for 10 minutes at room temperature and baked *in vaccuo* at 80°C for 2 hours.

A similar protocol with certain modifications was used for efficient transfer of DNA from CHEF gels to nylon membranes. Longer duration was required for all steps, which included depurination by 2 treatments of 15 minute each in 0.25 M HCl, a 45 minute treatment in denaturing solution, and a 45 minute treatment in neutralizing solution, followed by blotting transfer for 24 to 40 hours.

2.5.2 Dot blotting

A 96-well dot blotting vacuum system (Hybri · Dot Manifold; Bethesda Research Laboratories Inc.) was used for transferring samples of DNA onto nylon membranes. DNA in solution was denatured by mixing with an equal volume of 0.4 M NaOH for 30 minutes at room temperature before loading on membranes.

2.5.3 Colony Lifts

A sterilize nylon membrane was placed on the surface of an agar plate. Colonies were patched on the membrane using toothpicks and then the plate was incubated at 37°C for 3 to 10 hours. In order to lyse bacterial cells and denature DNA, the membrane was placed colony side uppermost on a series of solution-saturated 3MM paper filters: 10% (w/v) SDS for 3 minutes, denaturation buffer (same as the one in Section 2.5.1) for 7 minutes, neutralization buffer (same as the one in Section 2.5.1) for 4 minutes, twice. The membrane was then vigorously washed in 2 × SSPE three times to remove debris, air dried for 30 minutes, followed by baking at 80°C for 2 hours to fix the DNA to the membrane.

2.5.4 Preparation of DNA Probes

Radioactively labeled DNA probes used in hybridization analysis were prepared by two protocols: end labeling and random primer labeling.

The end labeling method was used for labeling short oligonucleotides that were synthesized by the MBSU. A typical 5'-end labeling reaction includes the oligonucleotide probe, γ-^{32}P-ATP (2 pmoles per pmole of DNA), spermidine solution (7.25 mM spermidine, 50 mM Tris-HCl (pH 8.0), 5 mM EDTA), T4 polynucleotide kinase (PNK) (Roche), and kinase buffer. After incubation at 37°C for 30 minutes, the enzyme was inactivated by incubation at 65°C for 10 minutes.

Long DNA probes were labeled using random primers and mixture of all possible hexameric oligonucleotides. Samples of DNA in solution were denatured by heating to 95°C for 2 minutes and cooled immediately on ice. The denatured DNA was then mixed with reaction buffer (50 mM Trsi-HCl (pH 8.0), 5 mM $MgCl_2$, 2 mM DTT, 200 mM HEPES, 0.4 μg/μl BSA (pH 6.6) as final concentrations), random hexamers (50 A_{260}units/ml, 2 μl), α-^{32}P-dCTP (1 μl), three dNTPs (dATP, dTTP and dGTP, final concentration of 1 mM for each) and 2 units of Klenow fragment of *E. coli* DNA polymerase I in a total volume of 20 μl. Reactions were incubated at room temperature (21 to 24°C) overnight or at 37°C for 2 hours, then terminated by heating at 95°C for 5 minutes, which also denatured the DNA, followed by immediate cooling on ice. DNA samples in low-melting point agarose gels were also labeled by random primer techniques, using a larger sample volume of DNA (30 to 35 μl) in a total reaction volume of 50 μl.

2.5.5 Hybridization and Autoradiography

Nylon membranes carrying transferred DNA were prepared for hybridization by incubation in hybridization buffer (6 × SSPE, 5 × Denhardt solution, 0.5% SDS, 50% formamide; Sambrook, *et al.*, 1989) with salmon sperm DNA added to a final concentration of 100 μg/ml, at an appropriate temperature (typically 42-45°C) for 2 hours. The denatured, labeled DNA probes were then added. Hybridization was carried out overnight at the same temperature. After hybridization, the membranes were washed twice with a low stringency solution (2 × SSPE and 0.2% (w/v) SDS) at room temperature for 5 to 10 minutes each, followed by two to three washes with a high stringency solution (0.1 × SSPE and 0.2% SDS) at 65°C for 30 minutes each, and they were finally rinsed with 0.1 × SSPE briefly at room temperature. The membranes were wrapped with Saran Wrap and exposed to X-ray film (Scientific Imaging Systems, Eastman Kodark Company) for an appropriate time, after which the films were developed in a film processor. An alternative method made use of a PhosphorImager^{TM} SI System (Amersham Biosciences) for autoradiography. Membranes were placed in the PhosphorImager^{TM} storage cassette and exposed to the phosphor screen for 2 to 8 hours. After exposure, the screen was scanned by the PhosphorImager (445 SI) to check results. The phosphor screen was erased by exposing it to visible light on a standard laboratory light box.

2.5.6 Stripping of DNA from Membranes

Radioactive probes from previously probed nylon membranes were removed by washing with 0.2 M NaOH at 42°C to 45°C twice, 10 minutes each time, followed by washing with 2 × SSPE at room temperature.

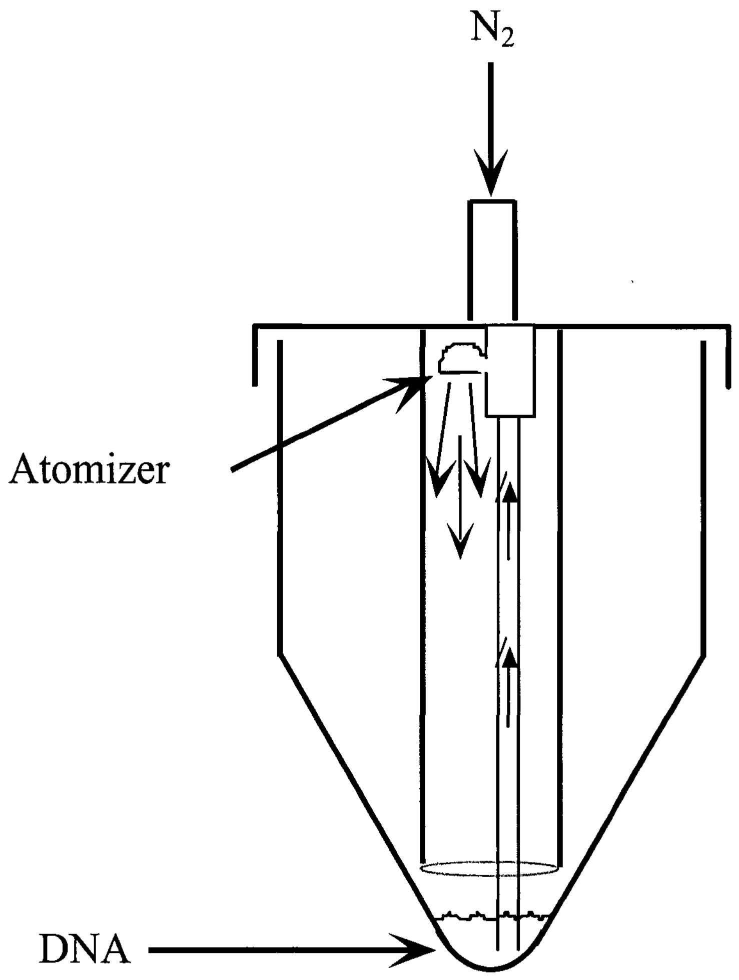
2.6 LIBRARY CONSTRUCTION

2.6.1 Random Fragment Libraries

pSCL2 DNA was dissolved in 750 μl shearing solution (Tris-EDTA, pH 8.0, containing 10% glycerol), and sheared into 1 to 3 kb or 3 to 6 kb fragments using a nebulizer (Invitrogen; Fig. 2-1) with an air pressure of 9 psi or 5 psi for 90 seconds. The sheared DNA was precipitated by mixing with an equal volume of 100% isopropanol and sit on dry ice for 15 minutes. The ends of the sheared DNA were made blunt by treating with T4 DNA polymerase and Klenow DNA polymerase at room temperature (21-25°C) for 30 minutes. The blunt-end repairing reaction included 35 μl sheared DNA in deionized water, 5 μl 10× blunting buffer (to a final concentration of 10 mM Tris-HCl, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM dithiothreitol, pH 7.9 at 25°C), 1 μl BSA (to a final concentration of 0.1 mg/ml), 5 μl dNTP mix (to a final concentration of 250 μM for

Fig. 2-1. Diagram of nebulizer used to shear DNA.

The various components are indicated by appropriate labels. This diagram was adapted from the literature supplied by Invitrogen.


N_2
Atomizer
DNA

each), 2 μl T4 DNA polymerase (4 U/μl in 20 mM potassium phosphate, pH 6.5, 5 mM dithiothreitol, 50% glycerol) and 2 μl Klenow DNA polymerase (4 U/μl in 50 mM potassium phosphate, pH 7.0, 0.25 mM dithiothreitol, 50% glycerol) in a total volume of 50 μl. After the reaction the enzymes were inactivated by incubation at 75°C for 20 minutes. For DNA dephosphorylation, 40 μl water, 5 μl 10× dephosphorylation buffer (to a final concentration of 25 mM Tris-HCl, 0.05 mM EDTA, pH 8.5 at 20°C), and 5 μl calf intestinal phosphatase (CIP; 1 U/μl in 25 mM Tris-HCl, 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$, 50% glycerol, pH 7.6 at 4°C) were added into the 50 μl blunt-end repair reaction in a total volume of 100 μl. The dephosphorylation reaction was incubated at 37°C for 60 minutes, then extracted with 50 μl phenol/chloroform, and precipitated by mixing with three volumes of cold ethanol. The blunt-end, dephosphorylated DNA was cloned into the pCR®4Blunt-TOPO® vector by mixing 4 μl DNA with 1 μl vector and 1 μl salt solution (1.2 M NaCl and 60 mM $MgCl_2$) provided by the TOPO® Shotgun Subcloning Kit obtained from Invitrogen. The T3 and T7 universal priming sites that flanked the cloning site in the vector could be used for sequencing the insert DNA on both strands. The ligated DNA was transformed into TOP10 One Shot® Electrocomp™ *E. coli* cells by electroporation. The transformed cells were regenerated in SOC medium for 30 minutes at 37°C, and then selected by spreading on LB plates containing 50 μg/ml kanamycin.

2.6.2 Nested Deletion Libraries

A nested deletion library was created using the Erase-a-base Kit (Promega). The DNA fragment of interest was first cloned into either a pGEM7Zf(+) or pGEM11Zf(-) vector (Promega) depending on the restriction sites on the insert DNA, so that at least two unique restriction sites lie between the end of the insert DNA and the sequencing priming site. The recombinant plasmids were then double digested with two different restriction enzymes to produce a 3'-overhang end and a 5'-overhang or blunt end adjacent to the end of the insert from which deletions are to proceed. Exo III enzyme (400 units) was added to each reaction at 42°C to obtain a deletion rate around 500 bp/min. Aliquots were removed from the reactions every 45 seconds. The ends were then repaired with Klenow

DNA polymerase and four dNTPs, then ligated with T4 DNA ligase. The recircularized plasmids were transformed into *E. coli* JM109 high efficiency chemically competent cells and selected by growth on LB plates containing 100 μg/ml ampicillin.

2.7 POLYMERASE CHAIN REACTION (PCR)

The primers used for PCR amplification were designed from the ends of assembled contigs using GeneTool software (BioTools Inc.). The lengths of the primers were between 18 and 22 nucleotides, and the melting temperature (T_m) of the primers were in the range of 58°C to 68°C, except for some rare cases. The melting temperature of primers with a length around 20 nucleotides can be calculated by the following formula:

$$T_m\ (^\circ\mathrm{C}) = (\text{number of G and C}) \times 4\ ^\circ\mathrm{C} + (\text{number of A and T}) \times 2\ ^\circ\mathrm{C}$$

The sequences and T_m of the PCR primers are listed in the Appendix.

A typical PCR reaction includes 1× PCR buffer (Roche), 5% DMSO, 0.2 mM of each dNTP (7-deaza dGTP partially replaced dGTP when it was necessary; Section 3.3.2.3; Fig. 3.3-1), 0.4 μM of each primer, up to 2 μg template DNA, *Taq* and *Pfu* polymerase in a total volume of 50 μl. A Genius FGEN02TP (Techne (Cambridge) Ltd.) or a MiniCycler^{TM} PTC-150 (MJ Research, Inc.) thermal cycler was used for PCR. The reaction conditions were as follows:

1 cycle	initial denaturation	96 °C	2 minutes
5 cycles	denaturation	96 °C	25 seconds
	annealing	T_m	25 seconds
	extension	72 °C	1 minute/kb
25 cycles	denaturation	96 °C	25 seconds
	annealing	T_m - 2 °C	25 seconds
	extension	72 °C	1 minute/kb + 5 sec/cycle
1 cycle	final extension	72 °C	15 minutes

2.8 DNA SEQUENCING

2.8.1 DNA Sequencing Reactions

The DYEnamic ET Terminator Cycle Sequencing Kit (Amersham Pharmacia Biotech.) was used for sequencing reactions. A standard sequencing reaction included 4 to 8 μl of sequencing reagent premix, 0.1 to 0.2 pmol of template DNA, and 5 pmol of primer, in a total volume of 20 μl. The template DNA could be either PCR products or plasmid DNA, which was purified using an RNase treatment, a phenol/chloroform (1:1) extraction and a PEG precipitation (equal volume of 13% PEG_{8000} with a final concentration of 0.5 M NaCl). If the template DNA contained a sequence that is very difficult to denature (such as a stretch of bases of G or C) or a complex secondary structure, 5% DMSO (with 0.8 to 1.5 M betaine (*N,N,N*-trimethylglycine) when it was necessary) was added in the 20 μl reaction mixture. A two-step, single primer amplification reaction was used to amplify the fluorescent signal when DNA samples with high GC content were sequenced using the T3 or T7 universal primers. Typical conditions included 30 cycles with a denaturation step of 96°C for 25 seconds, and a combined annealing and extension step at 58°C for 1.5 minutes. When other primers were used, the annealing and extension step was carried out at a temperature 2 to 5°C lower than the melting temperature (T_m) of the primer. All of the primers used for sequencing are listed in Appendix. PCR amplification was done with either a Genius FGEN02TP (Techne (Cambridge) Ltd.) or a MiniCycler^{TM} PTC-150 (MJ Research, Inc.) thermal cyclers. When cycling was complete, 2 μl (1/10 volume) of sodium acetate/EDTA buffer and 80 μl of 95% cold ethanol were added to each reaction and mixed well. DNA was allowed to precipitate for 15 minutes on ice, followed by centrifugation at 4°C for 15 minutes. The precipitated DNA was washed with 70% ethanol.

2.8.2 Automated DNA Sequencing

Double-stranded DNA sequencing was performed with an ABI Prism 377 or a 373 stretch with XL upgrade Automated Sequencer (Applied Biosystems), which was operated by Patricia Murray and Lisa Ostafichuk in the MBSU in the Dept. of Biological Sciences, University of Alberta. The chromatograms were viewed using the Edit View program (Applied Biosystems), GeneTool (BioTools Inc.) and Sequencher (Demo version; Gene Codes Corp.).

2.9 COMPUTATIONAL ANALYSIS

The raw sequences obtained by automated sequencing were edited using the EditSeq program and assembled with the SeqMan program in the DNAstar (Lasergene) Sequence Analysis package (DNASTAR, Inc.), as well as the Sequencher software (Demo version; Gene Codes Corp.). The DNAstar Sequence Analysis program package was also used for sequence alignment, protein secondary structure prediction and other analyses. Other computer programs used for DNA and protein sequence analysis included DNA Strider (Marck, 1988), GeneTool and PepTool (BioTools Inc.). The applications of these software packages are summarized in Table 2-1.

The program FramePlot (Bibb *et al.*, 1984; Ishikawa *et al.*, 1999) was used for the identification of open reading frames (ORFs). DNA and protein sequences were compared with those in GenBank at the NCBI (National Center for Biotechnology Information, Washington, DC, USA) using PSI-BLAST programs (Altschul *et al.*, 1997). InterProScan was used to identify to which known protein family (if any) a new sequence belongs to by scanning InterPro database. InterPro database consists of a group of member databases, including SWISS-PROT, PROSITE, Pfam, PRINTS, ProDom, SMART and TIGRFAMs. An online program, 3D-PSSM (3 Dimension - Position Specific Scoring Matrix), was used to predict the tertiary structure of the putative proteins. All of the online analysis tools used in this study are listed in Table 2-1.

Table 2-1. List of the software and online analysis tools used in this study.

Name	Function	Company or Website	Reference
Edit View	Chromatograms view	Applied Biosystems	
GeneTool	Chromatograms view Sequence editing, assembly, alignment Primer design, etc.	BioTools	
PepTool	Protein secondary structure prediction	BioTools	
DNAstar (Lasergene)	Sequence editing, assembly, alignment Protein secondary structure prediction Restriction map drawing, etc	DNASTAR, Inc.	
Sequencher	Sequence assembly, Chromatograms view	Gene Codes Corp.	
DNA Strider	Sequence editing, translating Restriction mapping Repeat sequence searching, etc		Marck, 1988
CLUSTAL W	Multiple sequence alignment	http://www.ebi.ac.uk/clustalw/#	
FramePlot	Open Reading Frames identification	http://watson.nih.go.jp/~jun/cgi-bin/frameplot-3.0b.pl	Ishikawa *et al.*, 1999
PSI-BLAST	Compare input protein sequences against other protein sequences in the GenBank	http://www.ncbi.nlm.nih.gov/BLAST/	Altschul *et al.*, 1997
BLAST 2 SEQUENCES	Alignment for two sequences	http://www.ncbi.nlm.nih.gov/BLAST/	
3D-PSSM	Protein tertiary structure prediction	http://www.sbg.bio.ic.ac.uk/servers/3dpssm/	Kelley, *et al.*, 2000
InterProScan	Protein classification	http://www.ebi.ac.uk/interpro/scan.html	
SMART	Simple Modular Architecture Research Tool	http://smart.embl-heidelberg.de	Schultz *et al.*, 1998

CHAPTER 3 RESULTS

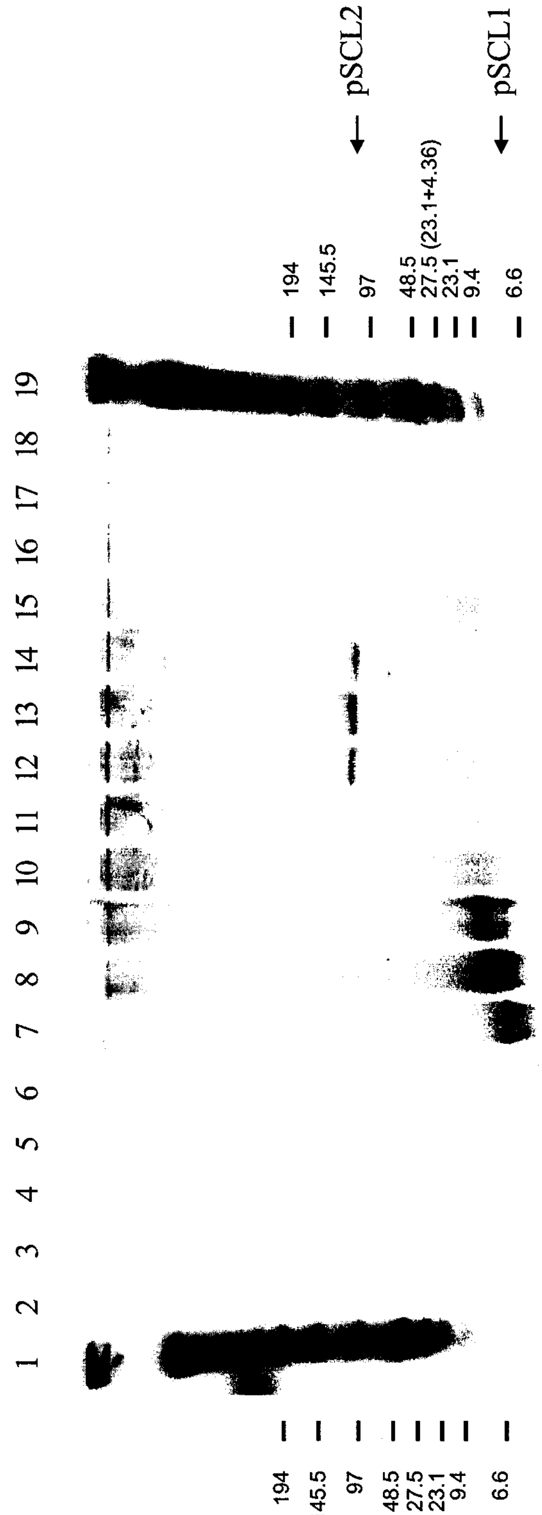
The two primary objectives of this research are determination of the nucleotide sequence of pSCL2, the 120 kb linear plasmid of *Streptomyces clavuligerus*, and annotation of the genes encoded therein. This chapter summarizes the experimental work involved and the results obtained in each stage of the research. This research started with DNA isolation and preliminary investigation of the restriction map of the linear plasmid. In the sequencing project, random fragment libraries were constructed, and DNA sequence was obtained, assembled and analyzed using various strategies. Based on the sequence information, the plasmid was annotated. An examination of the putative proteins that have essential functions or may play important roles in this linear plasmid is presented individually at the end of this chapter.

3.1 DNA ISOLATION

To start determination of the nucleotide sequence of pSCL2, it was essential to obtain a usable amount of pure plasmid DNA. Since the giant (larger than 50kb) linear plasmids are easily sheared in solution, and are not topologically constrained, they could not be purified by conventional methods, such as ethidium bromide/CsCl gradient sedimentation. The sucrose gradient centrifugation method was found to be able to isolate both small linear plasmids, such as pSCL1 (11.7 kb), and the middle-sized linear plasmids, such as pSCL2 (120 kb), from the genomic DNA in an intact form. Fig. 3.1-1 depicts the result of a typical sucrose gradient separation of *S. clavuligerus* DNA. Purified pSCL1 and pSCL2 are clearly visible as bands in lane 7 and lanes 12 to 14, respectively. However the small amount of pSCL2 plasmid DNA obtained from this technique was not enough for the subsequent required manipulations. pSCL3, the large linear plasmid, is too big to be isolated by this method, likely due to mechanical shearing during the procedure.

Fig. 3.1-1. Sucrose gradient purification of linear plasmid DNA from *S. clavuligerus*.

High molecular weight DNA was isolated from *S. clavuligerus* using proteinase K digestion in the presence of SDS (see Section 2.2.2.2 in Materials and Methods). This DNA was fractionated on a 20 to 50% sucrose gradient by centrifugation at 20,000 rpm for 65 hours in a SW40Ti rotor at 4°C. The gradient was fractionated and DNA in samples of each fraction was precipitated in an equal volume of 2-propanol for 12 hours at -20°C, centrifuged, and the supernatant discarded. The pelleted DNA was then redissolved and electrophoresed on a CHEF gel for 28 hours at 170 V with a 5 to 80 second pulse ramp at 12-15°C in 0.5 x TEB buffer containing 60 μM thiourea. The 1.5% agarose gel was prepared in 0.5 × TBE buffer with 12 mM thiourea. The gel was stained with 0.001% ethidium bromide after electrophoresis. Lanes 1 and 19 contain a low range PFG marker (lambda ladder + lambda DNA digested with *Hind*III); lanes 2 to 17 correspond to sucrose gradient fractions, with lane 2 coming from the top of the gradient, and lane 18 the bottom. Numbers to the side of the figure indicate molecular weight marker positions (length in kb).


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
194
145.5
97
48.5
27.5 (23.1+4.36)
23.1
9.4
6.6
← pSCL2
← pSCL1
194
145.5
97
48.5
27.5
23.1
9.4
6.6

In order to obtain larger amounts of giant linear plasmids from *S. clavuligerus*, immobilized cell lysis techniques and CHEF gel electrophoresis were applied. Cells embedded in agarose gels were lysed *in situ* with lysozyme, followed by digestion with proteinase K in the presence of SDS. After these treatments, only cellular DNA was left in the cavities remaining where the cells were originally located. The blocks containing the intact genomic DNA were loaded into wells and electrophoresed on a CHEF gel to separate the giant linear plasmids, pSCL2 and pSCL3, from the chromosomal DNA. Fig. 3.1-2 shows a typical separation of *S. clavuligerus* DNA molecules on a CHEF gel. The bands containing pSCL2 and pSCL3 were excised from the gel for further manipulation, such as *in situ* restriction enzyme digestion or DNA recovery by electroelution or β-agarase digestion.

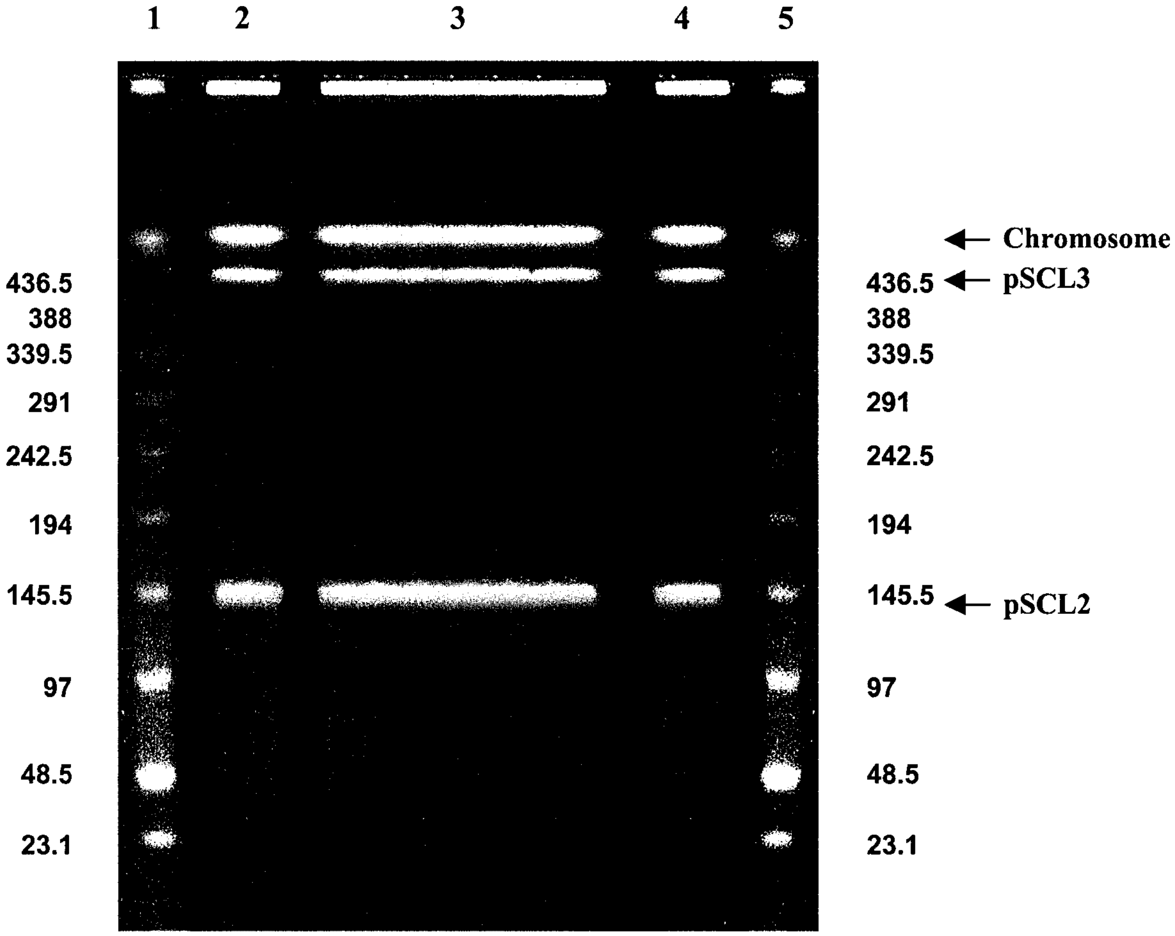
3.2 RESTRICTION ANALYSIS OF pSCL2 AND pSCL3

Because *Streptomyces* DNA has a high GC content, it has often been observed that restriction endonucleases that recognize A/T rich sites cut *Streptomyces* DNA at much lower frequencies than those enzymes that recognize G/C rich sites. Based on such observations, 24 restriction enzymes that recognized A/T rich sites were tested for the ability to digest pSCL2 and pSCL3. Single or double digested *S. clavuligerus* genomic DNA was examined using both CHEF and vertical conventional gel electrophoresis. pSCL2 or pSCL3 DNA, obtained from *in situ* cell lysis and CHEF gel electrophoresis, was labeled as a probe to screen Southern transfers of the total DNA digests. Previous studies showing no obvious homology between the plasmids and the chromosome in *S. clavuligerus* ensure the reliability of this strategy.

The preliminary survey of restriction sites of the pSCL2 and pSCL3 plasmids showed that the restriction enzyme *Ase*I, *BstZ17*I, *Dra*I, *Pme*I, *Ssp*I and *Swa*I did not produce any apparent digestion products from pSCL2, nor did *Dra*I, *Pme*I and *Swa*I digest pSCL3.

Fig. 3.1-2. CHEF gel electrophoresis of *S. clavuligerus* total DNA.

DNA was prepared from *S. clavuligerus* cells embedded in agarose blocks using the whole cell *in situ* lysis technique (see Section 2.2.2.1 in Materials and Methods). A 1.2 % agarose gel was prepared in 0.5 × TBE buffer containing 12 mM thiourea. The CHEF gel was run for 38 hours at 170 V, with a 3 to 40 second pulse ramp at 12 to 15°C in 0.5 × TEB buffer containing 60 μM thiourea. The gel was stained with ethidium bromide after electrophoresis. Lanes 1 and 5 contain a Low Range PFG marker (lambda ladder + lambda DNA digested with *Hind*III). Lanes 2 to 4 contain *S. clavuligerus* chromosome and the giant linear plasmids, pSCL2 and pSCL3. The small linear plasmid pSCL1 had run out of the gel. Numbers to the side of the figure indicate the positions of molecular weight markers (lengths in kb).


1
2
3
4
5
Chromosome
436.5
pSCL3
388
339.5
291
242.5
194
145.5
pSCL2
97
48.5
23.1

DNA fragments produced from a number of single and double restriction enzyme digestions of pSCL2 are summarized in Table 3.2-1. The restriction enzymes *Spe*I, *Hind*III, *Eco*RV, and *Eco*RI, making the fewest cuts (3, 5, 7, and 7, respectively) on pSCL2, were used to carry out double digestions. *Afl*II, which makes two cuts on the plasmid, was not selected for double digestion because of its unstable performance during digestion. The fragment numbers of some double digest were fewer than they should have been because some small fragments were not detectable. Some of these small fragments that could not be observed on either CHEF or conventional electrophoresis gels were deduced from sequence information and are marked in Table 3.2-1. Southern hybridization of some of these digestions is shown in Fig. 3.2-1 and Fig. 3.2-2. Determination of the location of the restriction sites within the plasmid proved not to be simple. A restriction map (Fig. 3.3-3) could only be constructed with the aid of hybridization experiments (Table 3.3-1) that will be discussed in section 3.3.2.4.

Similar restriction analysis was done for pSCL3. Single and double digests of pSCL3 using some rare-cutting restriction enzymes, *Ase*I, *BstZ17*I, *Hind*III, *Spe*I and *Ssp*I, are summarized in Table 3.2-2. *Ase*I, *Spe*I, *Ssp*I and *BstZ17*I, producing 2, 3, 4 and 8 fragments from pSCL3 respectively, were used to carry out double digestions. Because the sizes of all of the restriction fragments of pSCL3 were obtained from CHEF gels, fragments smaller than 20 kb may not have been detected. This could also be the reason that the number of fragments obtained from some double digests was fewer than expected. Southern hybridization of pSCL3 double digests is shown in Fig. 3.2-3. The location of the restriction sites of *Ase*I, *Spe*I and part of *BstZ17*I is illustrated in Fig. 3.2-4. The location of the remaining *BstZ17*I sites and *Ssp*I sites cannot be determined due to insufficient information.

The sizes of pSCL2 and pSCL3 were confirmed by adding up the sizes of all the fragments from various restriction digestions. The average size of pSCL2 is 120.5 kb, which is consistent with the results from the previous study (Netolitzky *et al.*, 1995). However, the average size of pSCL3, 464 kb, is 8% larger than the previous estimate, 430 kb (Netolitzky *et al.*, 1995).

Table 3.2-1. Summary of restriction fragments produced by digestion of pSCL2.

Restriction Enzyme(s)	Recognition Site	Number of Fragments	Fragments from CHEF (kb)	Fragments from Vertical Conventional Gel (kb)	Total Size (kb)
*Eco*RI	G/AATTC	8	30	22, 19, 13.5, 11.5, 10, 7.5, 4.5	118
*Eco*RV	GAT/ATC	8	50	20, 14.5, 13, 10.5, 7.3, 2.7, 1*	119
*Hind*III	A/AGCTT	7	39, 28.5	19, 11.8 (2[a]), 5.3, 2.3	117.7
*Spe*I	A/CTAGT	4	77, 29, 11, 6.5		123.5
*Eco*RI + *Hind*III	G/AATTC A/AGCTT	12		22, 19.5 (2), 13.5, 11.5, 10.5, 7.5, 5, 4.5, 3.1, 2.1, 1.9*	120.6
*Eco*RV + *Hind*III	GAT/ATC A/AGCTT	13		20, 19, 14.5, 13, 11.8, 10.5, 8.7, 7.3, 5.2, 3, 2.7, 2.3, 1*	119
*Eco*RI + *Eco*RV	G/AATTC GAT/ATC	11	27, 22, 19, 12, 11, 10, 8.5, 7, 4.5	2*, 1*	124
*Bcl*I	T/GATCA	29		10.8, 8.5, 8 (2), 6.8 (2), 6 (2), 5.5 (2), 4.5 (2), 4.3,4, 3, 2.7 (3), 2.6 (3), 2.4, 2, 1.9, 1.7, 1.5, 1.3, 1, 0.9	120.8
*Bgl*II	A/GATCT	15		21, 16.2, 11 (2), 10, 9, 7.1, 6 (3), 4.5, 4, 2.8, 1.8, 0.9	117.3
*Afl*II	C/TTAAG	3	100	19, 9	128
*Apo*I	R/AATTY	8		23, 20.4, 17.1, 16.3, 15.3, 10.6, 9.9, 5.2	117.8

* Fragments deduced from sequence information

a. The numbers in parentheses represent the numbers of fragments with similar sizes, which are listed in front of the parentheses.

Fig. 3.2-1. Southern hybridization of restriction endonuclease digests of *S. clavuligerus* genomic DNA fractionated using CHEF gel electrophoresis and probed with pSCL2.

A. Total DNA was prepared by *in situ* lysis of *S. clavuligerus* cells embedded in agarose. The DNA was digested *in situ* using a variety of restriction endonucleases. CHEF gel electrophoresis was carried out on a 1.2% agarose gel for 22.5 hours at 170 V using a 2 to 5 second ramp.

Marker lanes: 4 (M1): Low Range PFG marker (λ ladder + λ *Hind*III digest);
8 (M2): *Xho*I and *Bau36*I double digested λ DNA.

Sample lanes: 1 (EI): *Eco*RI digest;
2 (EI EV): *Eco*RI and *Eco*RV double digest;
3 (EV): *Eco*RV digest;
5 (H EI): *Hind*III and *Eco*RI double digest;
6 (H EV): *Hind*III and *Eco*RV double digest;
7 (H): *Hind*III digest;
9 (S): *Spe*I digest;
10 (U): undigested *S. clavuligerus* DNA .

The numbers to the side of the figure indicate the positions of molecular weight markers (lengths in kb).

B. Southern hybridization of the gel in panel A probed with labeled pSCL2. Hybridization was carried out at 43°C in the presence of 50% formamide (see Section 2.5 for more detail).

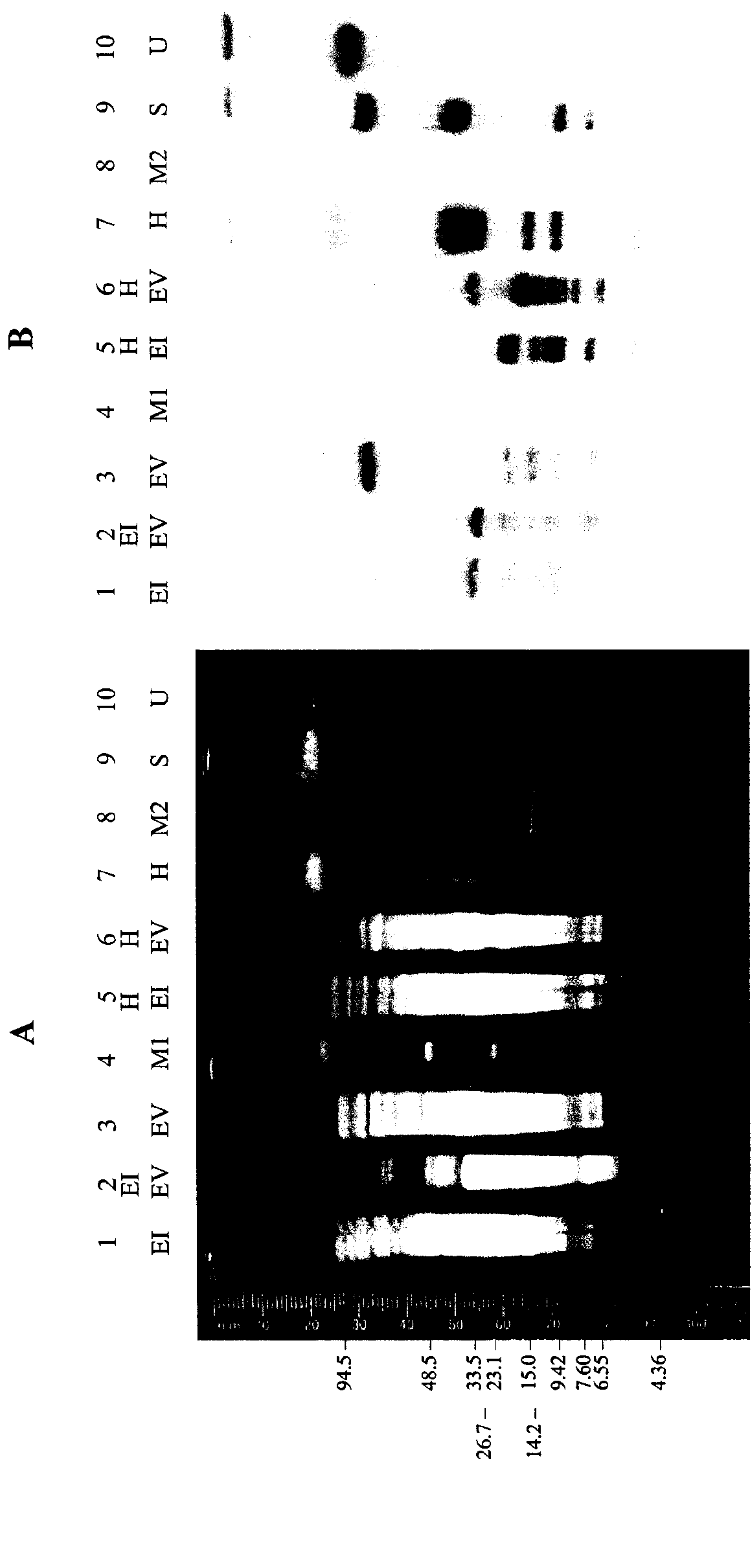

A
1 EI
2 EI EV
3 EV
4 M1
5 H EI
6 H EV
7 H
8 M2
9 S
10 U
94.5
48.5
33.5
26.7
23.1
15.0
14.2
9.42
7.60
6.55
4.36
B
1 EI
2 EI EV
3 EV
4 M1
5 H EI
6 H EV
7 H
8 M2
9 S
10 U

Fig. 3.2-2. Southern hybridization of restriction endonuclease digests of *S. clavuligerus* genomic DNA fractionated using conventional agarose gel electrophoresis and probed with pSCL2.

A. *S. clavuligerus* DNA, prepared by a quick method for total DNA isolation, was digested using a variety of restriction endonucleases and fractionated on a conventional 0.5% agarose gel for 16 hours at 2 V/cm.

Marker lanes: 5 (M1): *Hind*III and *Bau36*I double digested λ DNA;
8 (M2): *Eag*I and *BstE*II double digested λ DNA.

Sample lanes: 1 (U): undigested *S. clavuligerus* DNA;
2 (H): *Hind*III digest;
3 (H EV): *Hind*III and *Eco*RV double digest;
4 (EV): *Eco*RV digest;
6 (H EI): *Hind*III and *Eco*RI double digest;
7 (EI) *Eco*RI digest;
9 (BII): *Bgl*II digest;
10 (BI): *Bcl*I digest.

The numbers to the side of the figure indicate the positions of molecular weight markers (lengths in kb).

B. Southern hybridization of the gel in panel A probed with labeled pSCL2. Hybridization was carried out at 43°C in the presence of 50% formamide (see Section 2.5 for more detail).

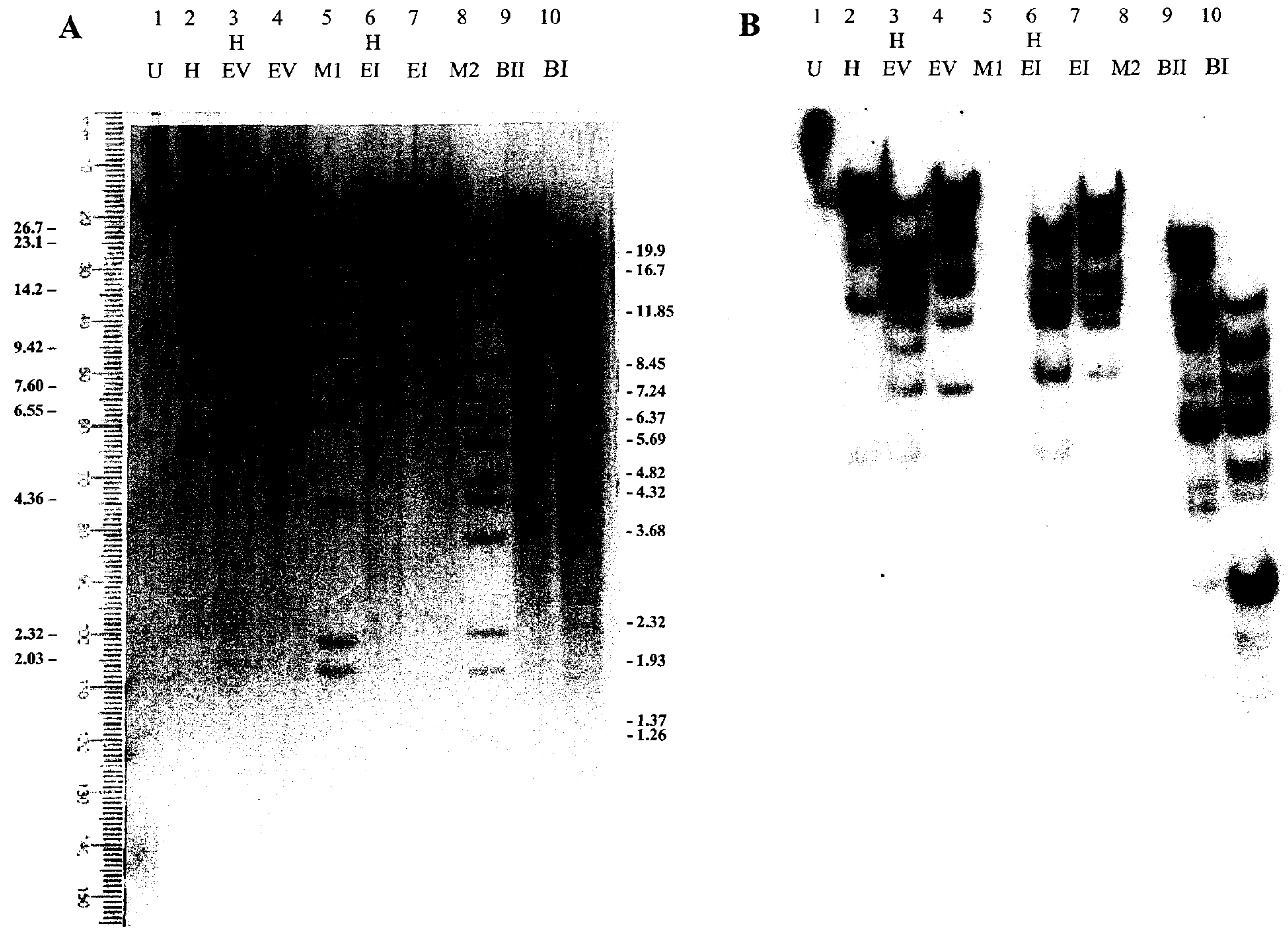

A
1 2 3 4 5 6 7 8 9 10
H H
U H EV EV M1 EI EI M2 BII BI
26.7 –
23.1 –
14.2 –
9.42 –
7.60 –
6.55 –
4.36 –
2.32 –
2.03 –
- 19.9
- 16.7
- 11.85
- 8.45
- 7.24
- 6.37
- 5.69
- 4.82
- 4.32
- 3.68
- 2.32
- 1.93
- 1.37
- 1.26
B
1 2 3 4 5 6 7 8 9 10
H H
U H EV EV M1 EI EI M2 BII BI

Table 3.2-2. Summary of restriction fragments produced by digestion of pSCL3.

Restriction Enzyme(s)	Recognition Site	Number of Fragments	Fragments from CHEF (kb)	Total size (kb)
*Ase*I	AT/TAAT	2	255, 210	465
*Spe*I	A/CTAGT	3	275, 105, 75	455
*BstZ17*I	GTA/TAC	8	160, 100, 80, 55, 20, 19, 16, 16	466
*Ase*I + *Spe*I	AT/TAAT A/CTAGT	4	255, 105, 75, 35	470
*Ase*I + *BstZ17*I	AT/TAAT GTA/TAC	9	165, 100, 60, 50, 25, 18, 18, 15, 15	466
*BstZ17*I + *Spe*I	GTA/TAC A/CTAGT	9	100, 80, 80, 75, 60, 23, 21, 20, 20	479
*Ase*I + *Ssp*I *	AT/TAAT AAT/ATT	5	200, 160, 60, 20, 20	460
*Ssp*I *	AAT/ATT	4	220, 160, 65, 20	465
*Hind*III *	A/AGCTT	7	145, 110, 50, 50, 40, 40, 18	453

* May be digested incompletely

Fig. 3.2-3. Southern hybridization of restriction endonuclease digests of *S. clavuligerus* genomic DNA fractionated using CHEF gel electrophoresis and probed with pSCL3.

A. Total DNA was prepared by *in situ* lysis of *S. clavuligerus* cells embedded in agarose. The DNA was digested *in situ* using a variety of restriction endonucleases. CHEF gel electrophoresis was carried out on a 1.2% agarose gel for 36 hours at 170 V using a 3 to 40 second ramp time.

Marker lanes: 7 (M): Low Range PFG marker (λ ladder + λ *Hind*III digest);

Sample lanes: 1 (A + B): *Ase*I and *BstZ17*I double digest;
2 (A + H): *Ase*I and *Hind*III double digest;
3 (A + Sp): *Ase*I and *Spe*I double digest;
5 (A + Ss): *Ase*I and *Ssp*I double digest;
6 (B + Ss): *BstZ17*I and *Ssp*I double digest;

The numbers to the side of the figure indicate the positions of molecular weight markers (lengths in kb).

B. Southern hybridization of the gel in panel A probed with labeled pSCL3. Hybridization was carried out at 43°C in the presence of 50% formamide (see Section 2.5 for more detail).

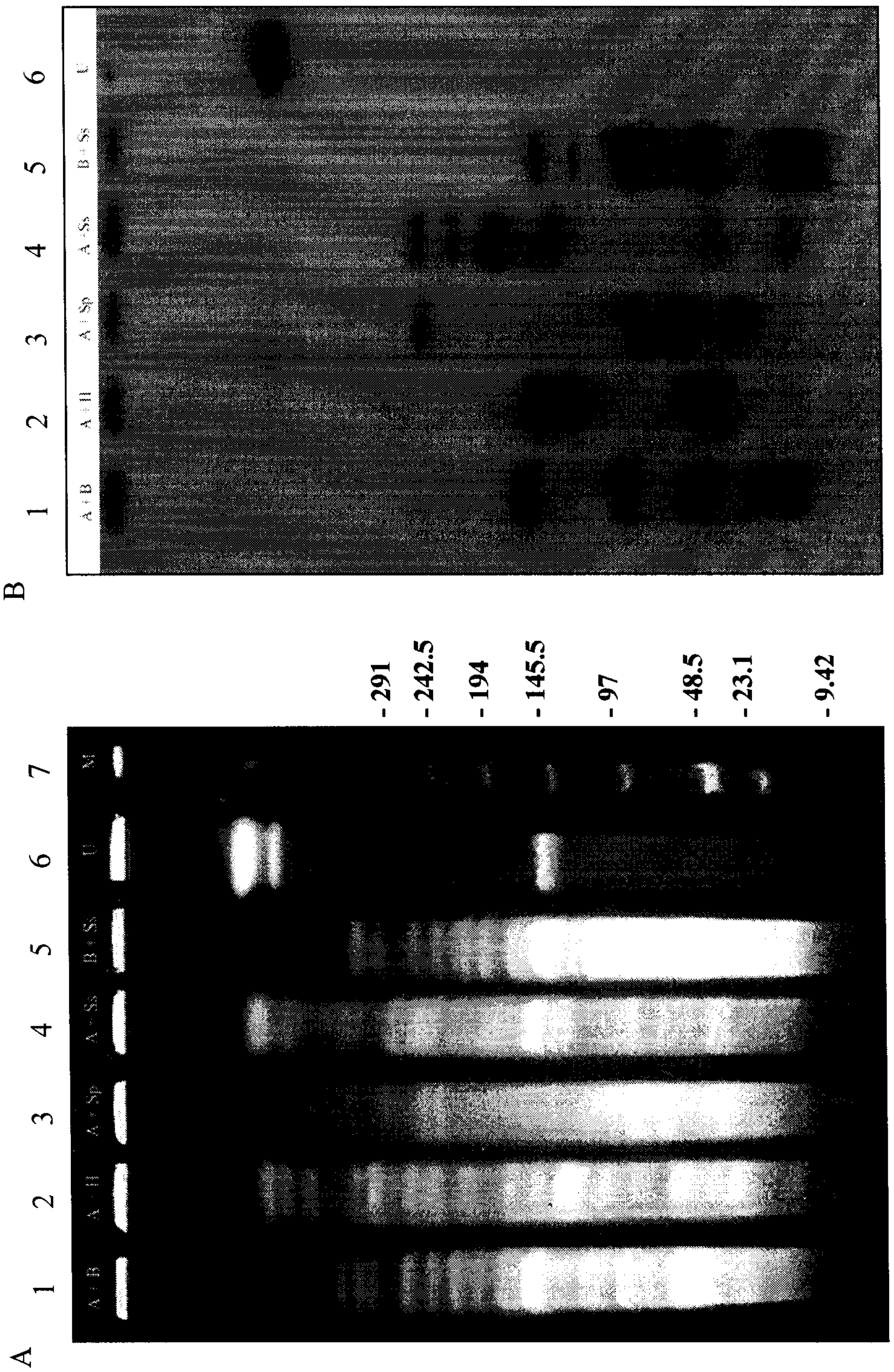
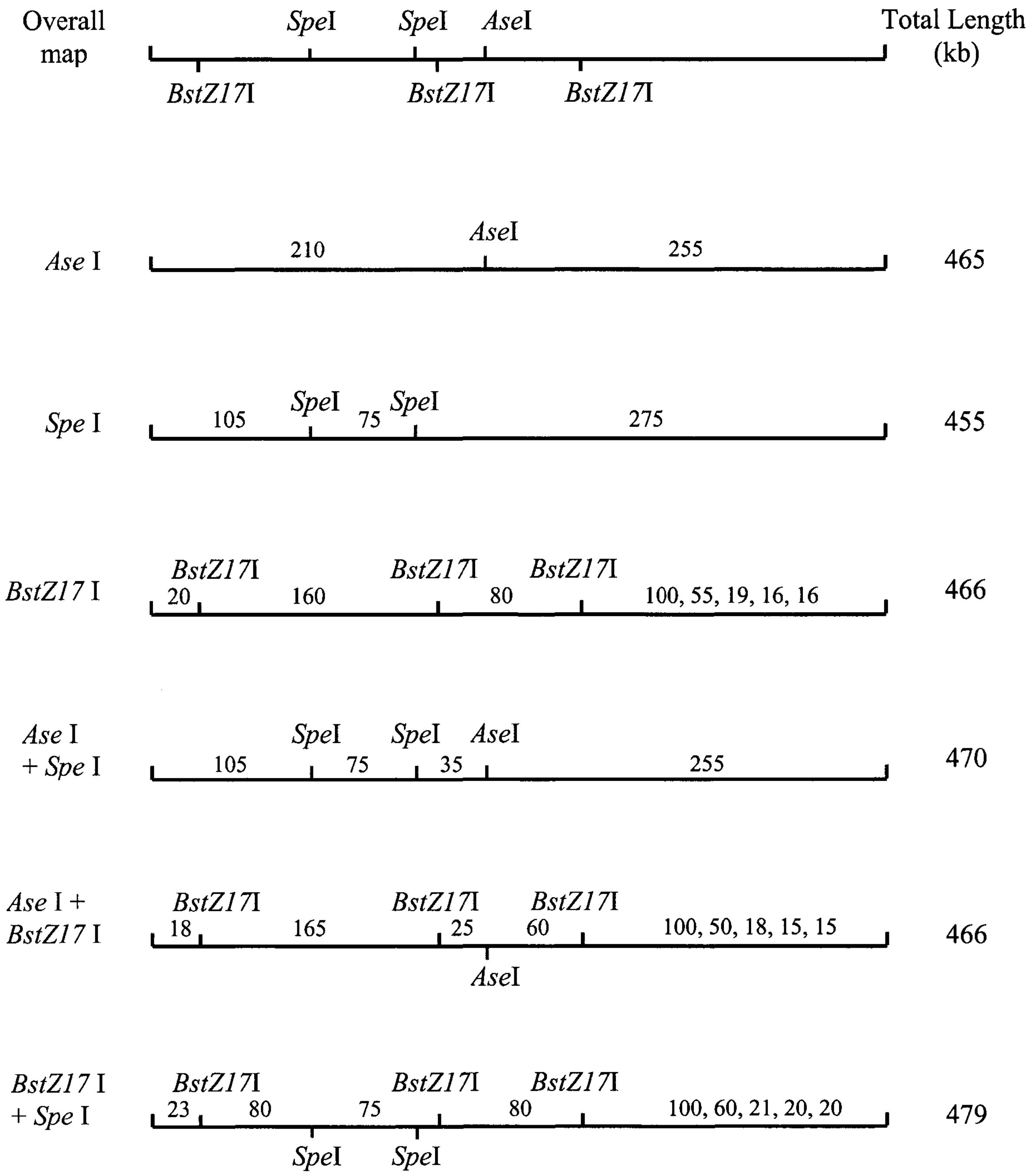

A
1
2
3
4
5
6
7
A+B
A+H
A+Sp
A+Ss
B+Ss
U
M
-291
-242.5
-194
-145.5
-97
-48.5
-23.1
-9.42
B
1
2
3
4
5
6
A+B
A+H
A+Sp
A+Ss
B+Ss
U

Fig. 3.2-4. Restriction map of pSCL3.

Restriction fragments were produced by the digestion reactions listed in Table 3.2-2. The overall map is shown at the top. The recognition sites of *Ase*I, *Spe*I and *BstZ17*I are represented by vertical stripes. The numbers indicate fragment lengths in kb. The total lengths of the plasmid, which were obtained by adding up the sizes of all the fragments from various restriction digestions, are also shown.


Overall map
SpeI
SpeI
AseI
Total Length (kb)
BstZ17I
BstZ17I
BstZ17I
Ase I
210
AseI
255
465
Spe I
105
SpeI
75
SpeI
275
455
BstZ17 I
BstZ17I
BstZ17I
BstZ17I
20
160
80
100, 55, 19, 16, 16
466
Ase I + Spe I
SpeI
SpeI
AseI
105
75
35
255
470
Ase I + BstZ17 I
BstZ17I
BstZ17I
BstZ17I
18
165
25
60
100, 50, 18, 15, 15
AseI
466
BstZ17 I + Spe I
BstZ17I
BstZ17I
BstZ17I
23
80
75
80
100, 60, 21, 20, 20
SpeI
SpeI
479

3.3 CLONING AND SEQUENCING STRATEGIES

The main strategies that have been used in a variety of genome-sequencing projects are divided into two groups: the ordered-clone approach and the random-sequencing (shotgun) approach. Both of these approaches were applied in this research. The random shotgun strategy proved to be a better one for the overall sequencing of pSCL2, whereas, primer-walking and nested-deletion that are both ordered approaches, were necessary in the later stages.

3.3.1 Ordered-clone Approach

Initially, an ordered subcloning approach was tried. It was planned that two libraries would be constructed: first one containing fragments with a size of 5 to 10 kb produced by enzymes that recognize A/T rich sites, such as *Bcl*I (T/GATCA) or *Bgl*II (A/GATCT); the other one containing approximately 1-kb fragments produced by frequently cutting enzymes, such as *Sma*I (CCC/GGG), or restriction enzymes recognizing 4-base-pair sites, such as *Rsa*I (GT/AC), *Alu*I (AG/CT) and *Mbo*I (/GATC). The small-insert clones in the second library would then be used in hybridization experiments to order the large-insert clones in the first library. Selected clones would then be sequenced by primer-walking or a nested deletion approach to obtain the complete sequences. However this plan was not successful because the restriction sites of these enzymes are not sufficiently evenly distributed on the plasmid. Smaller fragments were cloned much more easily than larger fragments. In order to obtain even-sized fragments, a random fragmentation method was found to be a better choice.

3.3.2 Random Fragmentation Approach

Shotgun sequencing is the most widely used strategy for a large sequencing project nowadays. Sequencing libraries of randomly sheared DNA can be constructed in a variety of cloning vectors. A large number of clones from libraries representative of the whole target DNA can be sequenced and assembled into contigs in the assembly phase. Ideally, the contigs can be joined together into one single contig in the closure phase if all the gaps can be filled using a variety of methods.

3.3.2.1 Library construction

The linear plasmid pSCL2 used for library construction was separated from the chromosomal DNA and the other two linear plasmids, pSCL1 and pSCL3, by CHEF electrophoresis and recovered by electroelution (Section 2.3.3.2). Using a nebulizer, samples of pSCL2 DNA were sheared into small fragments randomly, and these were cloned into the pCR®4Blunt-TOPO® vector (Section 2.6.1). Two libraries were constructed in this project: a smaller-insert (1 to 3 kb) and a larger-insert (3 to 6 kb) library. The small-insert library consisted of more than 500 clones, which provided approximately tenfold coverage of the complete plasmid. The larger-insert library was necessary to obtain a "scaffold" of the plasmid that was used during the closure phase.

During library construction it is important to keep a single DNA insert in each recombinant clone. The presence of multiple insertions in a given clone would cause assembly artifacts. The use of a TOPO cloning vector should eliminate any multiple fragment clones. Dephosphorylating the ends of the insert fragments before DNA cloning further avoided tandem inserts of two independently cloned fragments.

One problem encountered in library construction with this approach was that some regions of the pSCL2 sequence were not included in the libraries, though the total coverage (tenfold) of the libraries should have been high enough. On the contrary, it was also observed that some regions were cloned repeatedly. Several clones from two independent libraries had exactly the same inserts. The probable causes of these problems will be discussed in the next chapter.

3.3.2.2 Sequencing strategy

Most pSCL2 sequencing was performed using the standard sequencing reaction with double-stranded circular DNA templates (Section 2.8.1). Because *Streptomyces* DNA has an unusually high GC content, a two-step and single primer amplification reaction was used for DNA amplification. However, the sequencing of the extremely GC-rich regions often required further modification of the procedure. Addition of DMSO may decrease the extreme annealing temperature caused by high GC content (Sun *et al.*, 1993). Normally 5% DMSO was added to sequencing reactions, but this concentration may be varied in different cases. DMSO was also used together with 0.8 to 1.5 M betaine (*N,N,N*-trimethylglycine), which was reported to be able to reduce the formation of secondary structure caused by GC-rich sequences, either by binding to AT base pairs in the major groove (Rees *et al.*, 1993), or by increasing the hydration of GC pairs by binding within the minor groove (Mytelka & Chamberlin, 1996). Introducing nucleotide analogs, such as 7-deaza dGTP, into PCR reactions is a more direct way to decrease the secondary structure formed by GC content. However completely replacing dGTP with 7-deaza dGTP drastically reduces the yield of PCR products, thus lowering signals during automated sequencing. A replacement of 50 to 75% dGTP proved to be the best ratio to optimize the amplification in the tested cases (Fig. 3.3-1). Template denaturation by NaOH can also improve PCR amplification (Agarwal & Perl, 1993). Long DNA templates or DNA templates having high GC content were denatured by adding 0.4 M NaOH and 0.4 mM EDTA for 10 minutes at room temperature, followed by alcohol precipitation.

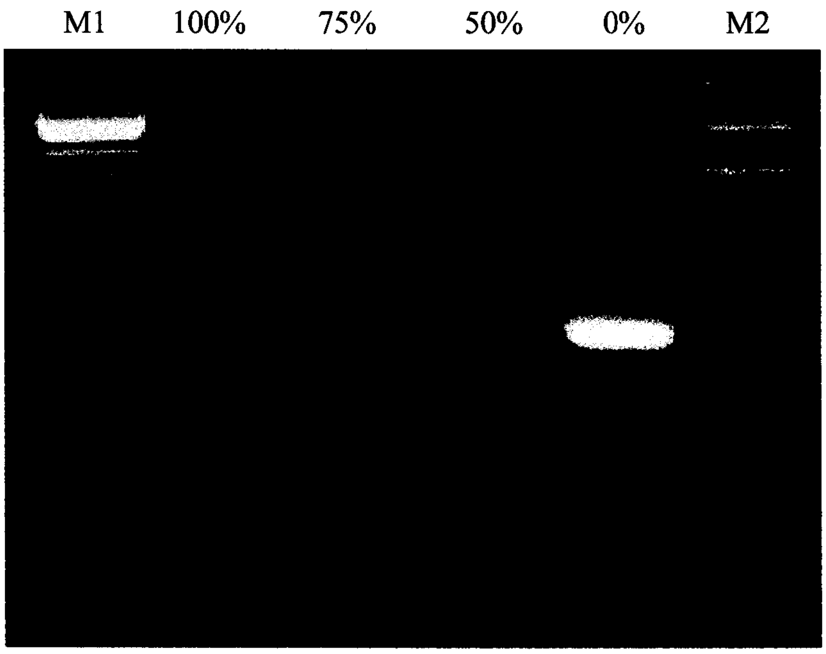
3.3.2.3 Assembly phase

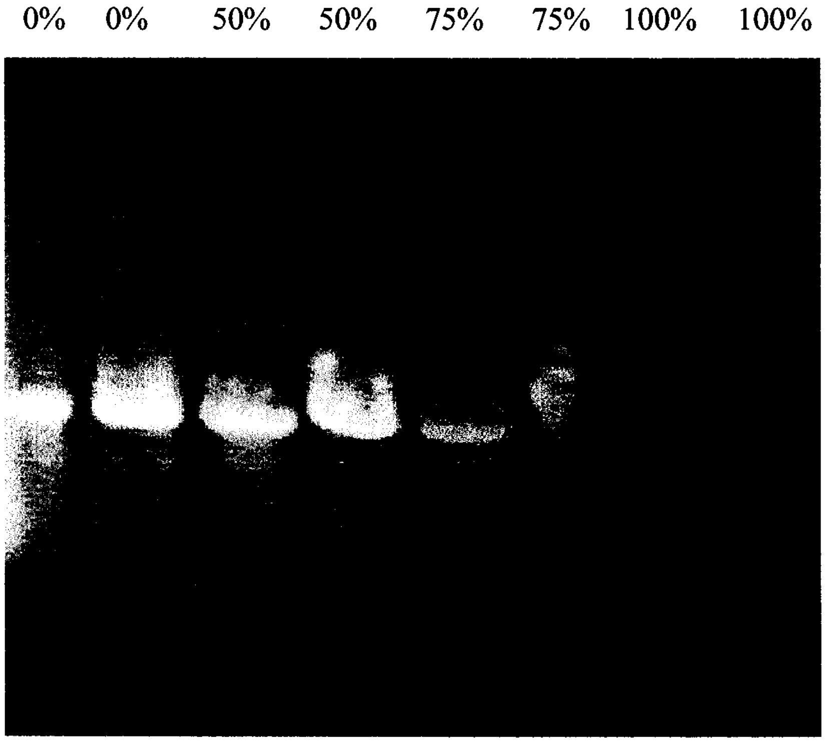
Raw sequences obtained from automated sequencing were edited and input into computer programs to be assembled. The various software tools used during these steps are listed in Table 2-1.

The contig number increased rapidly at a coverage level lower than one-fold since the sequences obtained at this stage statistically tend to be spread out instead of contiguous. In the second stage, the number of contigs decreased in proportion to the number of new sequences added to the assembly. The sequences obtained from new clones continue to spread out and extend the short contigs so that the gaps were filled and short contigs were linked into longer contigs. In the last stage, the decrease in the number of contigs gradually slowed down while the number of new sequences kept growing. More redundant sequences were obtained to increase the coverage of the sequence. The fewer the gaps that were left, the harder they were to be filled. This is an indication that one should shift from the random strategy to a more directed strategy for the closure phase.

Fig. 3.3-1. PCR amplification of pSCL2 DNA using varying concentrations of 7-deaza dGTP.

Standard PCR conditions were used (see Section 2.7 for the details) except that the nucleotide dGTP was replaced by 100%, 75%, 50% or 0% 7-deaza dGTP. The top and bottom panels demonstrate different yield and purity in each reaction, respectively. M1 and M2 in the top panel are molecular markers. Parallel results are shown in the bottom panel.


M1
100%
75%
50%
0%
M2


0%
0%
50%
50%
75%
75%
100%
100%

3.3.2.4 Closure phase

After contig assembly, the first step in the closure phase is to order the contigs. Several approaches were used to order contigs. A section that contains two or more putative neighboring contigs with small gaps (the lengths of the gaps were normally known) between them was defined as "gapped-contig".

The "primer-walking" method was the most frequently used method to extend contigs in this study. All the synthetic primers designed for this project are listed in the Appendix. The insert DNA of each clone is sequenced from both ends by the T3 and T7 universal primers. If the two terminal sequences of a single clone belong to different contigs, it is highly probable that these two contigs are adjacent to each other (Fig. 3.3-2A). Thus walking-primers were designed to anneal to the ends of the contigs to complete this gap. A large-insert clone is more likely to be a linking clone in this case. However, not all the gaps were covered by linking clones. Whatever the coverage of the genome, physical gaps may remain due to unclonable regions.

The order of the contigs provides important information for the subsequent steps. Hybridization is a basic method to order contigs. Clones at the ends of each contig, or clones containing useful restriction sites were labeled as probes to hybridize to restriction fragments of pSCL2, in the case of this study produced by *Spe*I, *Hind*III, *Eco*RV, and *Eco*RI. If the end sequences of two contigs hybridize to the same restriction fragment, the two contigs are very likely to be neighbors (Fig. 3.3-2B). On the other hand, if one clone hybridizes to two restriction fragments, these two fragments are likely to be adjacent, so that hybridization is also helpful to complete the restriction map of pSCL2 mentioned in the last section. The results of hybridization of different clones in each contig to the restriction fragments of pSCL2 are summarized in Table 3.3-1. Fig. 3.3-3 illustrates the order of ten contigs and the locations of *Spe*I, *Hind*III, *Eco*RV, and *Eco*RI restriction sites.

The contig order can also be deduced from the analysis of coding sequences (CDSs) predicted for each contig (Fig. 3.3-2C). If the comparison of the CDSs from contig ends to protein databases shows that the ends of two contigs encode different parts

Fig. 3.3-2. Approaches to link contigs.

A. The ends of two contigs belong to one linking clone.

B. Contigs can be ordered by hybridization. The "XXX" symbol represents positive hybridization.

C. The ends of two contigs show sequence homology to the same gene from another sequenced genome. The "|||||"symbol represents the similarity detected by BLAST.

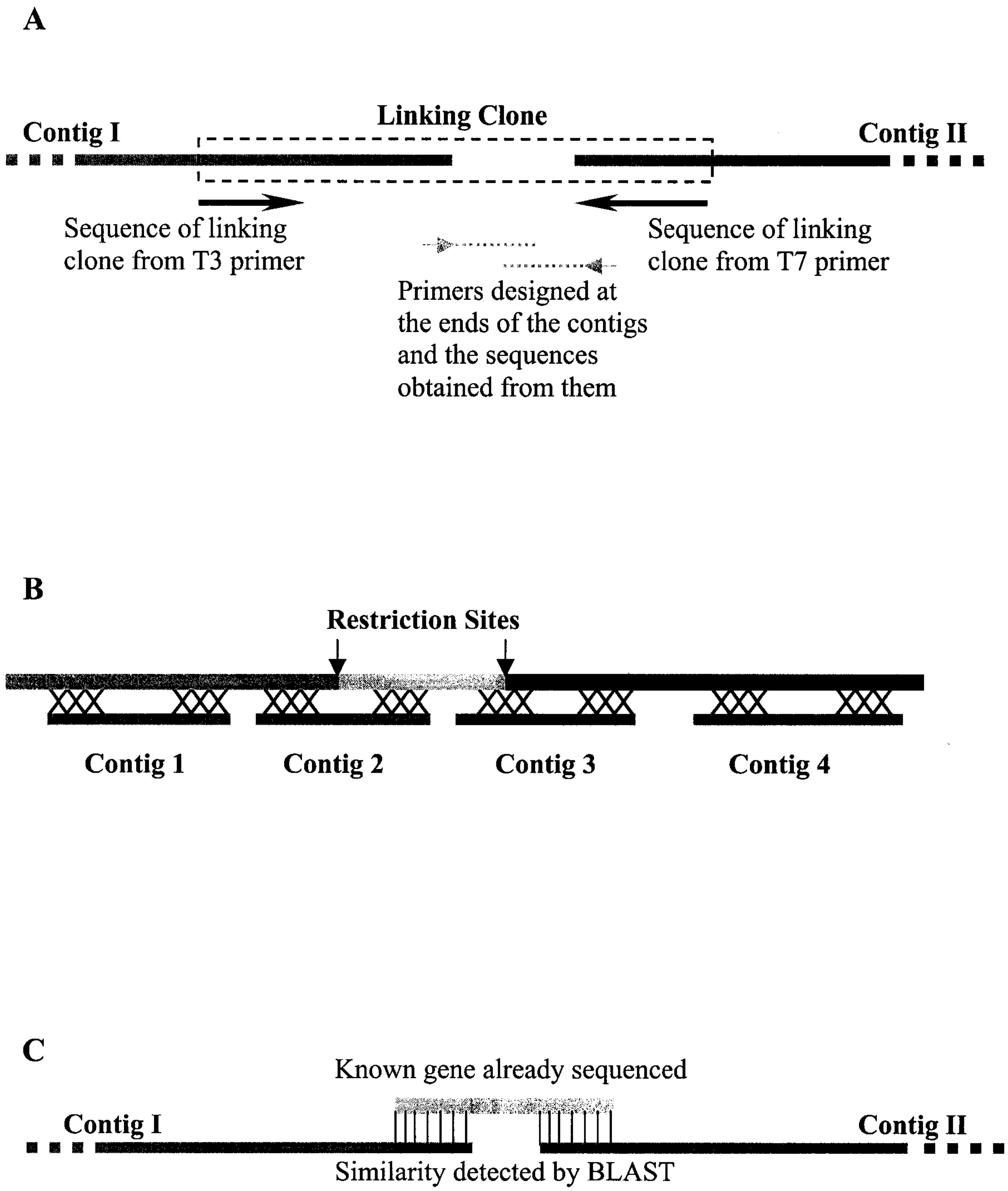

A
Linking Clone
Contig I
Contig II
Sequence of linking clone from T3 primer
Sequence of linking clone from T7 primer
Primers designed at the ends of the contigs and the sequences obtained from them
B
Restriction Sites
Contig 1
Contig 2
Contig 3
Contig 4
C
Known gene already sequenced
Contig I
Contig II
Similarity detected by BLAST

Table 3.3-1. Hybridization of *Spe*I, *Hind*III, *Eco*RV, and *Eco*RI fragments of pSCL2 to determine the order of the contigs.

+, positive hybridization; $+^1$, putative positive by mapping; $+^2$, putative positive to small fragments that can not be directly observed from electrophoresis gel, but verified by sequence information; $+^3$, false hybridization probably due to incomplete digestion.

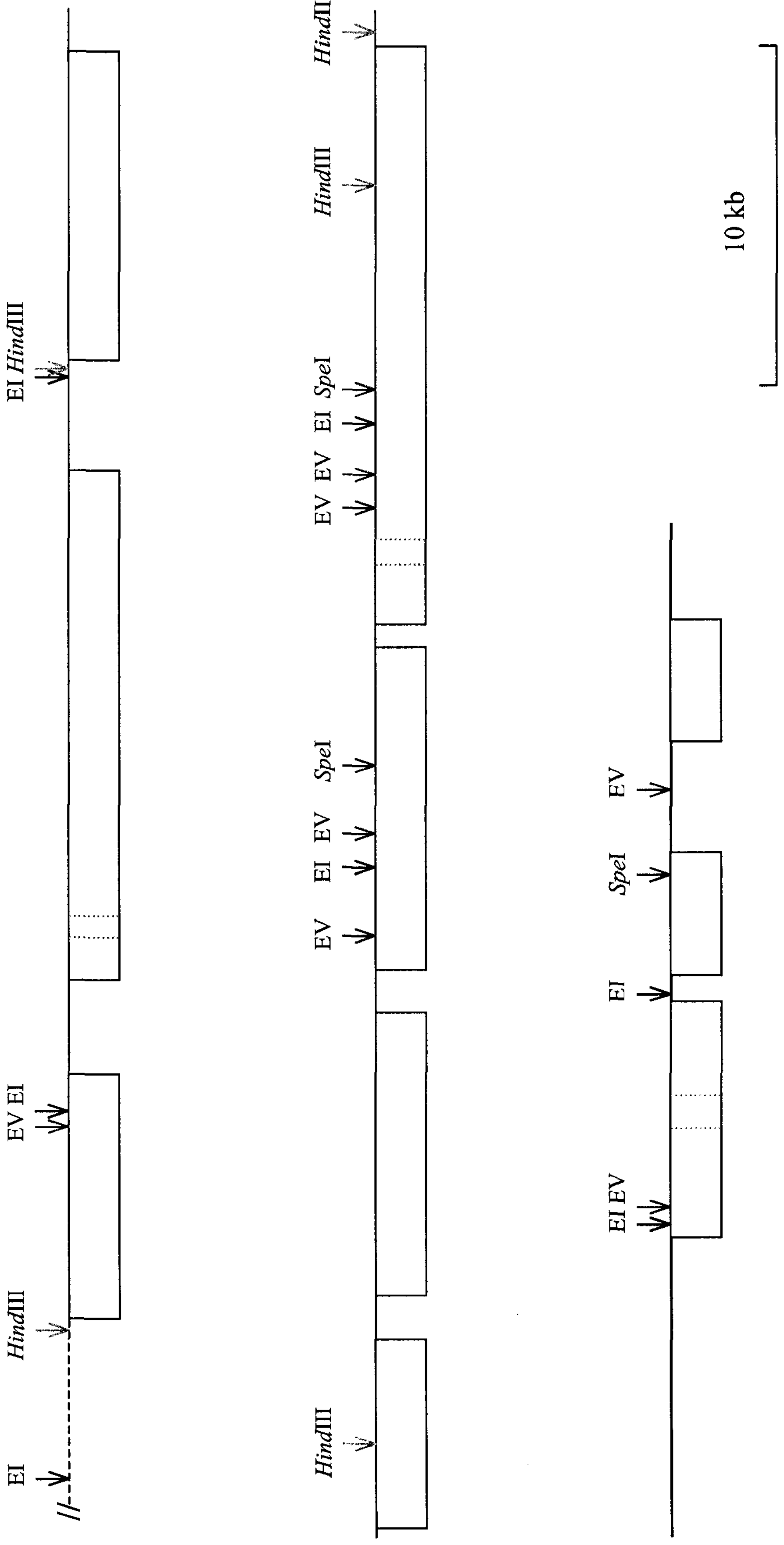
a. Clones used as probe.

b. L, left end of the contig; R, right end of the contig.

		Left End	Contig68		Contig26		Contig73		ContigH9		Contig 424	ContigA8		Contig207		Contig25		Contig69		Contig Ter
									L[b]	R[b]		L	R	L	R			L	R	
Fragments	Size (kb)		SG99[a]	SG224	SG188	SG67	SG172	SG111	SG80	SG274	SG429	SGA3	SG298	2ISG12	SG199	SGA5	SG171	SG69	SG169	Bc9
*Spe*I - A	77	+[1]	+	+	+	+	+	+	+	+	+	+								
C	11												+	+						
B	29													+	+	+	+	+	+	+[3]
D	6.5																		+	
*Hind*III-B	28.5		+	+	+	+														
D1	11.8						+	+	+	+										
A	39									+	+	+	+	+		+[3]	+[3]	+[3]	+[3]	+[3]
E	5.3														+					
C	19																			
D2	11.8	+[1]																		
F	2.3																			
*Eco*RV-A	50		+	+	+	+	+	+	+	+	+	+								
G	2.7											+[2]								
E	10.5												+							
H	1													+[2]						
B	20													+	+					
C	14.5															+	+	+	+	
F	7.3																			+
D	13	+[1]																		
*Eco*RI - E	11.5		+	+																
B	22			+	+	+														
A	30						+	+	+	+	+	+								
D	13.5												+	+						
C	19													+	+					
G	7.5															+	+			
F	10																	+	+	+
H	4.5	+[1]																		

Fig. 3.3-3. Order of the contigs (gapped-contigs) and the restriction map of pSCL2.

The broken line at the left end of the sequence is the area indicated as "Left End" in Table 3.3-1. The assembled contigs are shown as boxes below the axis, with their names labeled. Areas between the vertical dotted lines are gaps within the gapped-contigs. The recognition sites of *Spe*I, *Hind*III, *Eco*RI (EI) and *Eco*RV (EV) are shown above the axis. The lengths of the restriction fragments produced by single and double digestion are listed in Table 3.2-1.


EI
HindIII
EV EI
EI HindIII
HindIII
EV
EI
EV
SpeI
EV
EV
EI
SpeI
HindIII
HindIII
EI EV
EI
SpeI
EV
10 kb

of the same protein, it is probable that these two contigs are adjacent to each other. This method was used to order contig25 and contig69 in this study.

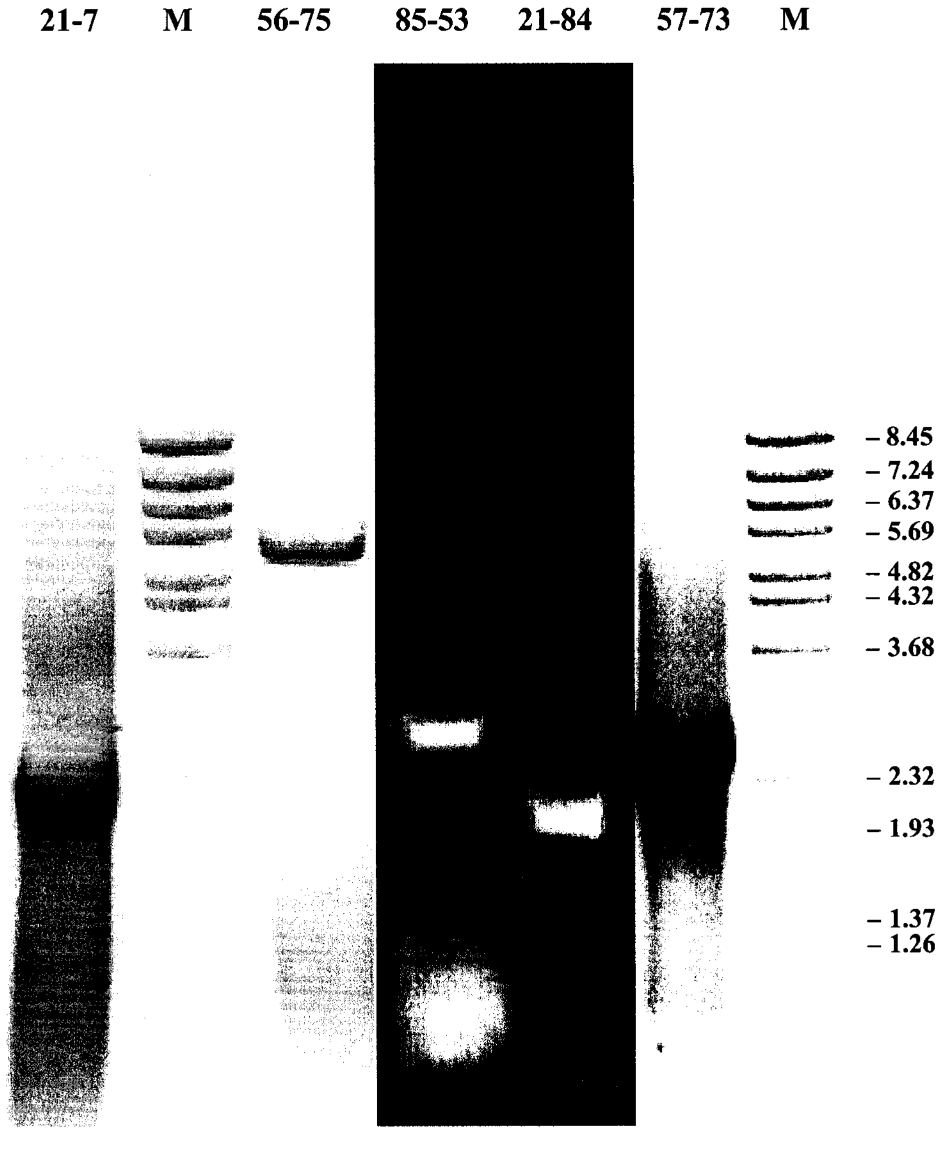
The prediction of contig order then needs to be verified by PCR amplification. Specific PCR products were amplified between two primers designed at the ends of putative neighboring contigs, using *S. clavuligerus* total DNA as template (Fig. 3.3-4). Thermostable *Taq* DNA polymerase is a very efficient polymerase that is widely used in general PCR amplification, which synthesizes DNA at a rate of more than 1 kb/min. However, it allows misincorporations due to the absence of proofreading capacity, which causes errors when amplifying DNA. Proofreading thermostable DNA polymerases, such as *Pfu*, possess an integral 3'-exonuclease activity to remove misincorporated nucleotides (Cheng *et al.*, 1994), but these DNA polymerases have a relatively low yield and lower amplification rate. The combination of *Taq* DNA polymerase and a lower level of *Pfu* DNA polymerase at a ratio of 10:1 was employed and shown to successfully amplify DNA fragments up to 5 kb in good yield. These amplified fragments not only provided the information on the gap size, but were also used as templates for primer-walking to fill gaps. However, sequencing reactions using PCR products as templates did not give satisfactory results that showed either short or noisy readings.

3.3.3 Nested Deletions

Some regions in the pSCL2 sequence contain abundant tandem repeats, which made automatic assembly very difficult. In order to solve this problem, the exonuclease III unidirectional nested-deletion approach (Henikoff, 1984; 1987) was applied to generate ordered overlapping DNA clones for sequencing. This method is based on two useful features of exonuclease III: processive digestion at a very uniform rate, and usually failure to initiate digestion at DNA ends with four-base 3'-overhanging ends. A large number of overlapping clones were quickly generated from certain pSCL2 fragments using this method. The main advantage of this technique is that the sequences obtained from the overlapping clones are ready to be assembled manually and allow nonrandom sequence analysis. This method is limited by the time required to obtain suitable clones in vectors with appropriate restriction endonuclease sites flanking the cloned DNA, and to

Fig. 3.3-4. PCR products designed to verify the order of contigs.

The numbers on each lane represent the primers used to amplify the DNA fragment loaded in the lane. For example, Product 85-53 resulted from amplification of the DNA located between primer 85 and primer 53. All of the primers and the contigs to which they belong to are listed in Appendix. Product 85-53 links contig68 and contig26; products 21-7 and 21-84 link contig73 and contigH9; product56-75 links contig424 and contigA8; product 57-73 links contigA8 and contig207. Lanes containing molecular weight markers are labeled as M and contain λ DNA digested with *BstE*II. Numbers to the side of the figure indicate the positions of marker (lengths in kb).


21-7
M
56-75
85-53
21-84
57-73
M
- 8.45
- 7.24
- 6.37
- 5.69
- 4.82
- 4.32
- 3.68
- 2.32
- 1.93
- 1.37
- 1.26

isolate the resulting deletion clones. Clones obtained after deletion experiments were examined by agarose gel electrophoresis, an example of which is shown in Fig. 3.3-5.

3.4 DNA SEQUENCE OF pSCL2

About 400 clones were isolated from the sequencing libraries. The insert DNA of each clone was sequenced from both ends using the T3 and T7 universal primers. Totally about 1000 sequence reactions were carried out. The average read length for each sequence reaction was about 600 bp. The raw sequence data was edited and assembled into contigs with the DNAstar Sequence Analysis package. Approximately 80% of the nucleotide sequence of the linear plasmid pSCL2 from *Streptomyces clavuligerus* has been completed with a total coverage of 8 to 9 folds. Both strands of the DNA were sequenced at least once; most regions were sequenced 2 to 5 times on each strand. 83 kb of the plasmid DNA was assembled into large contigs (more than 2 kb), including 7 continuous contigs and 3 gapped-contigs. The three gapped-contigs consist of two contigs and a small gap (less than 1 kb) in each of them (Fig. 3.3-3). The complete sequences of these contigs have been submitted to GenBank. Table 3.4-1 summarizes the length, GC content and the accession number of each contig. The largest contig has a length of 15 kb and the smallest one has a length of 2.1 kb. The overall GC content of the assembled sequence is 69.97%, comparable to that of other linear plasmids (71.9% in pSCL1 (*S. clavuligerus*), 68.4% in SLP2 (*S. lividans*), 69.1% in SCP1 (*S. coelicolor*), 69.2% in SAP1 (*S. avermitilis*), 70% in pSV2 (*S. violaceoruber*), and chromosomes (72.1% in *S. coelicolor*, and 70.7% in *S. avermitilis*) in *Streptomyces*.

3.5 GENE ANNOTATION

3.5.1 Open Reading Frame (ORF) Analysis

When a nucleotide sequence has been determined and assembled into contigs, one of the first objectives is to predict the presence of open reading frames (ORFs) that potentially encode proteins. With *Streptomyces* DNA sequences this is best accomplished

Fig. 3.3-5. DNA fragments produced by nested deletions, examined by agarose gel electrophoresis.

Linearized DNA (pSCL2 DNA cloned in the vector of pGEM7Z(+), the total length was approximately 7 kb) was digested with 400 units of exonuclease III at 42°C giving a deletion rate of approximately 500 bp/min. At each time point, (indicated above each lane), a 2.5 µl sample was taken from the digest reaction and mixed with 7.5 µl of S1 nuclease (50 units) for 30 minutes at room temperature. The S1 nuclease digestion was stopped by adding 1 µl of S1 stop buffer and inactivated by heating to 70°C for 10 minutes. A 2 µl portion was taken from each time point sample to determine the extent of digestion by electrophoresis on a 0.7% agarose gel. The marker lane is labeled as M and contained λ DNA digested with *BstE*II. Numbers to the side of the figure indicate the positions of marker (length in kb).

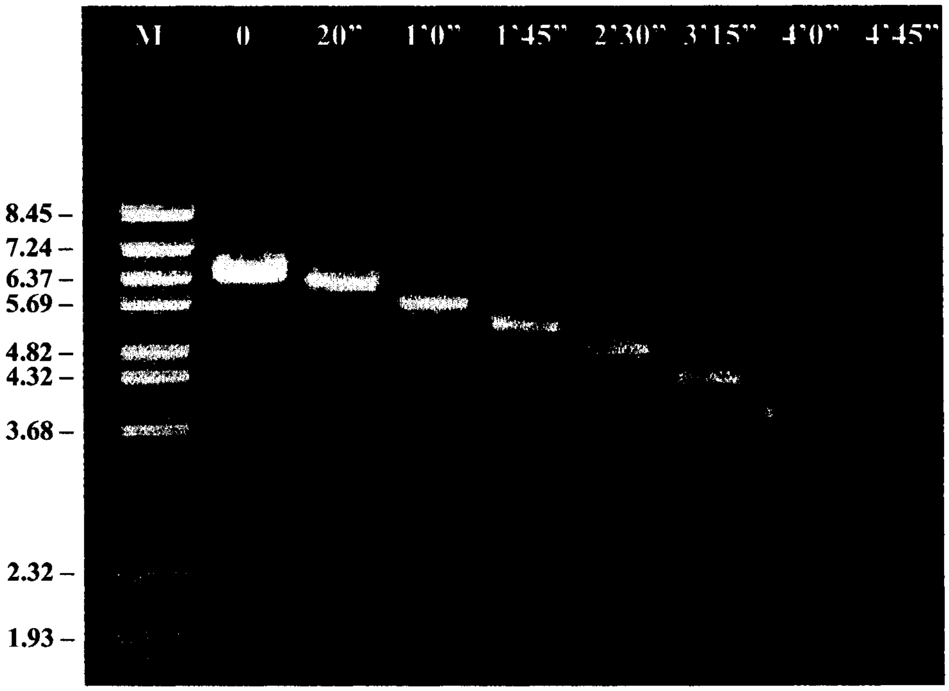

M
0
20"
1'0"
1'45"
2'30"
3'15"
4'0"
4'45"
8.45 –
7.24 –
6.37 –
5.69 –
4.82 –
4.32 –
3.68 –
2.32 –
1.93 –

Table 3.4-1. Summary of contigs (gapped-contigs) in pSCL2.

Contig Number	Length (bp)	GC Content (%)	Accession Number
1.68	7255	70.4	AY392410
*2.26-I	3569	71.4	AY392411
*2.26-II	10718	71	AY392412
3.73	9090	69.7	AY392413
4.H9	5599	70.9	AY392414
5.424	8391	70.4	AY392415
6.A8	9765	69.8	AY392416
*7.207-I	1656	68.8	AY392417
*7.207-II	15013	68.7	AY392418
*8.25-I	3224	66.5	AY392419
*8.25-II	2754	71.5	AY392420
9.69	3618	70.3	AY392421
10.Ter	2076	70.2	AY392422
Overall	82728	69.97	

* The shaded boxes represent the gapped-contigs, and the sub-contigs within each gapped-contig were named as I or II.

using the online program, FramePlot. ATG, GTG and TTG were set as potential initiation codons, while TAA, TGA and TAG were defined as terminator codons. Other parameters set in this program that may influence the output include Minimum ORF Size (10 or 20 codons), Window Size (40 codons) and Step Size (1 codon) for the resolution setting, and Incomplete ORF (ON) to allow detection of incomplete ORFs at the end of a contig. The graphical output of 10 contigs (or gapped-contigs) of pSCL2 is shown in Fig. 3.5-1. All possible ORFs in six frames (three in one direction and three in the other direction) were plotted as lines between the potential start and stop codons (here an ORF is defined as a nucleotide sequence that lacks potential translation stop codons; Kieser, *et al.*, 2000). Not surprisingly, numerous ORFs were predicted in all six possible frames throughout the entire sequence. In order to distinguish authentic protein-coding ORFs from all other possible ORFs, the high GC content of *Streptomyces* DNA is constructively useful. Genes of bacteria that have a high GC content DNA such as *Streptomyces* have a strong, biased codon usage (Bibb *et al.*, 1984). This results in uneven GC distribution at each position within the codons: over 90% GC at the third position, around 70% GC at the first position, and about 50% GC or even lower at the second position. Based on this observation, any ORF in which the GC percentages at the three positions within its codons show the characteristic bias was considered as a potential protein-coding ORF. Conversely, a region in which the GC content at each codon position is similar and around the average could be an intergenic region.

However, not all protein-coding ORFs have such an obvious, characteristic GC distribution. A possible ORF is considered as a putative protein-coding gene if it encodes a protein with a 40% or higher overall identity to a known gene in GeneBank. The regions shorter than 20 codons were not considered as a potential protein-coding ORF even if it contains a characteristic GC bias. The only exception is ORF 7.207.7 (13 codons) located immediately upstream of the replication proteins. The gene organization in this replication origin region in pSCL2 shows a similar pattern with that in the linear plasmid pSLA2-L in *Streptomyces rochei.* More detail will be discussed in Section 4.3.

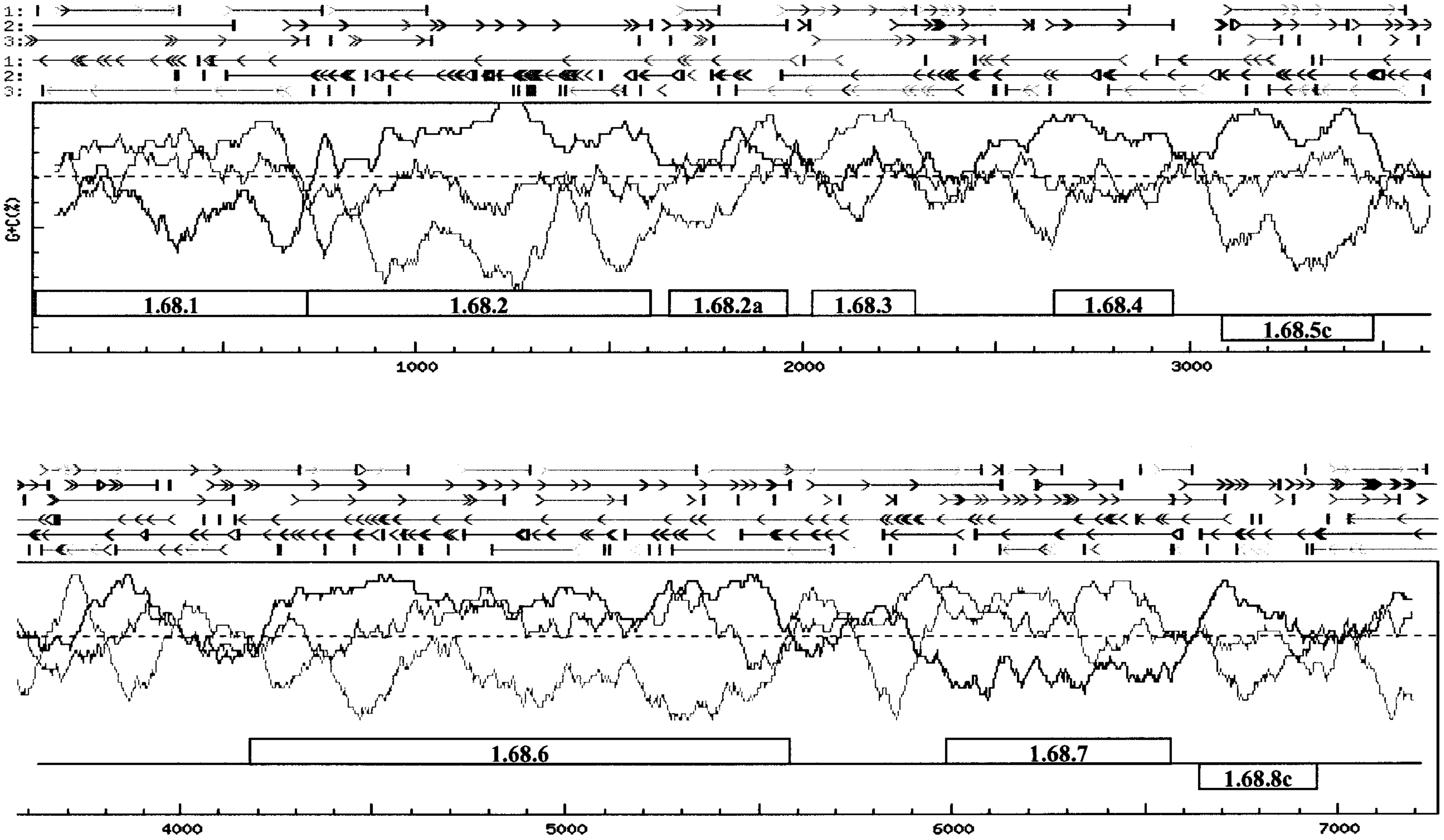
Potential ribosome binding sites (RBS) can also help to distinguish protein-coding ORFs. The typical RBS for translation initiation in *Streptomyces* is a 3-7 nt, A/G rich

Fig. 3.5-1. Frame analysis of the ten contigs determined in pSCL2.

The sequences of the ten contigs determined in pSCL2 were analyzed by the FramePlot program to reveal ORFs with the potential to encode proteins. Data of each frame are indicated by a different color. All predicted ORFs (three in each direction) are plotted above the frame plot as horizontal bars with potential start (>) and stop (|) codons. Putative protein-coding ORFs are shown under the plot, depicted as boxes placed either above (frames 1 to 3) or below (frames 4 to 6) the axis. The colors of the boxes indicate the frame that the putative protein-coding ORFs are located. The number of each ORF is labeled inside the box.

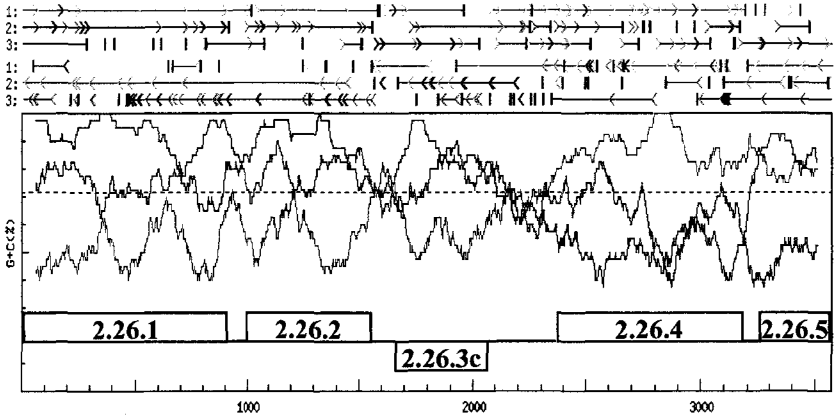
Contig 1.68

```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 68 7255 bp; 70.4% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 5 01:11:56 2003
```



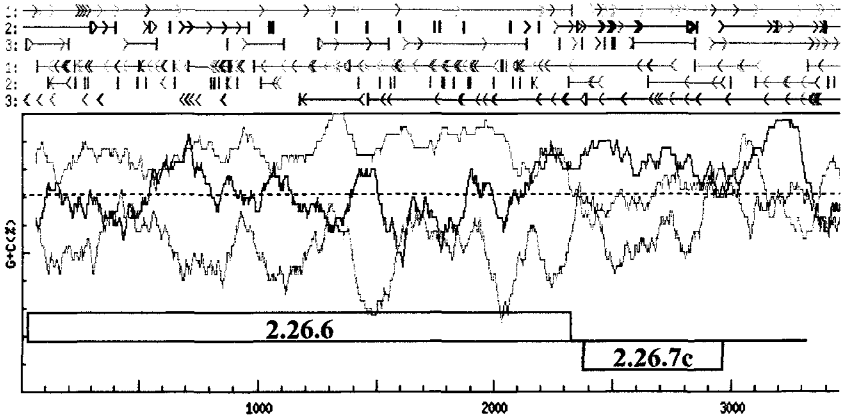
Contig 2.26-I

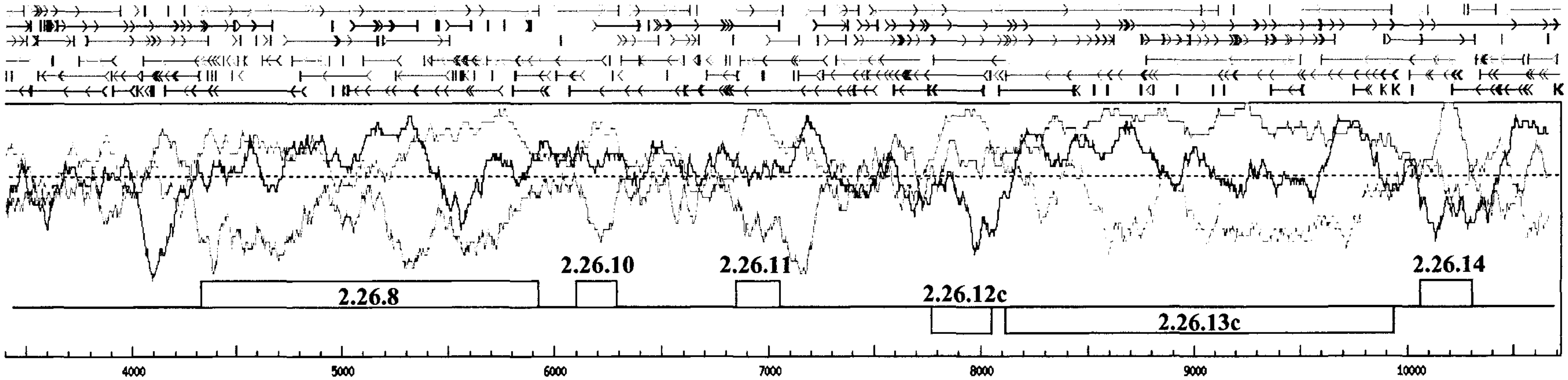
```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target:  3569 bp; 71.4% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 4 04:31:12 2003
```



Contig 2.26-II

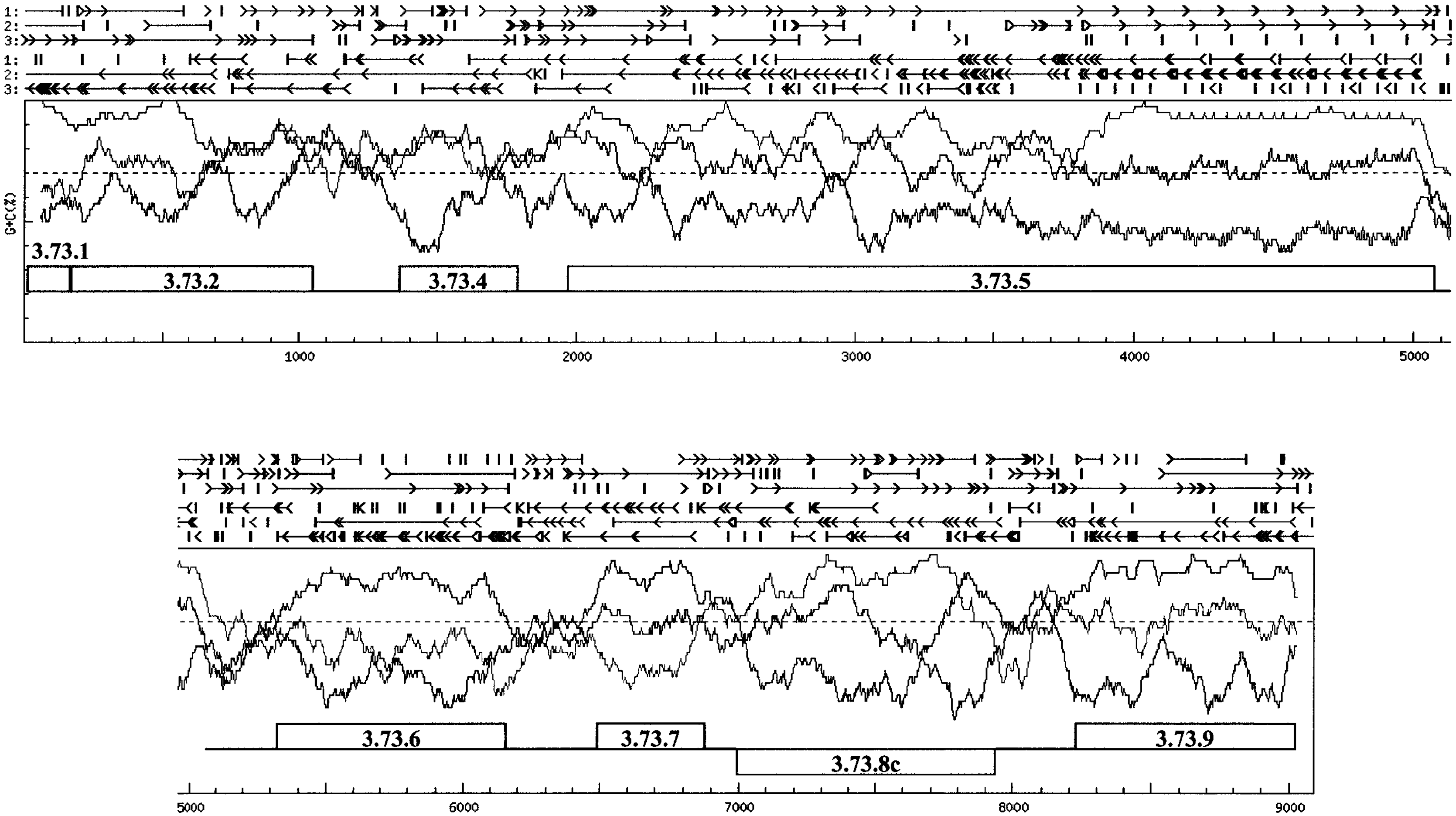
```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target:  10718 bp; 71.0% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 4 05:40:12 2003
```

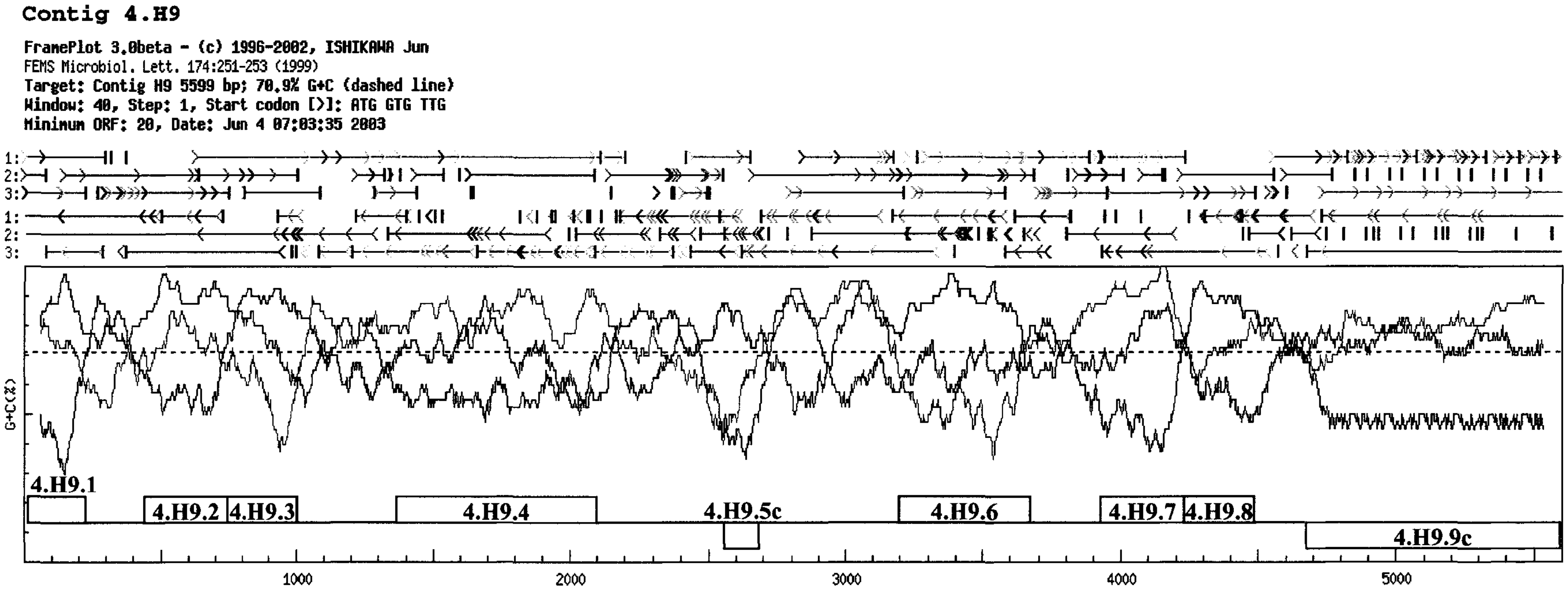




Contig 3.73

```
FramePlot 2.3.2 - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 73 9090 bp; 69.7% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Sep 19 05:07:31 2003
```



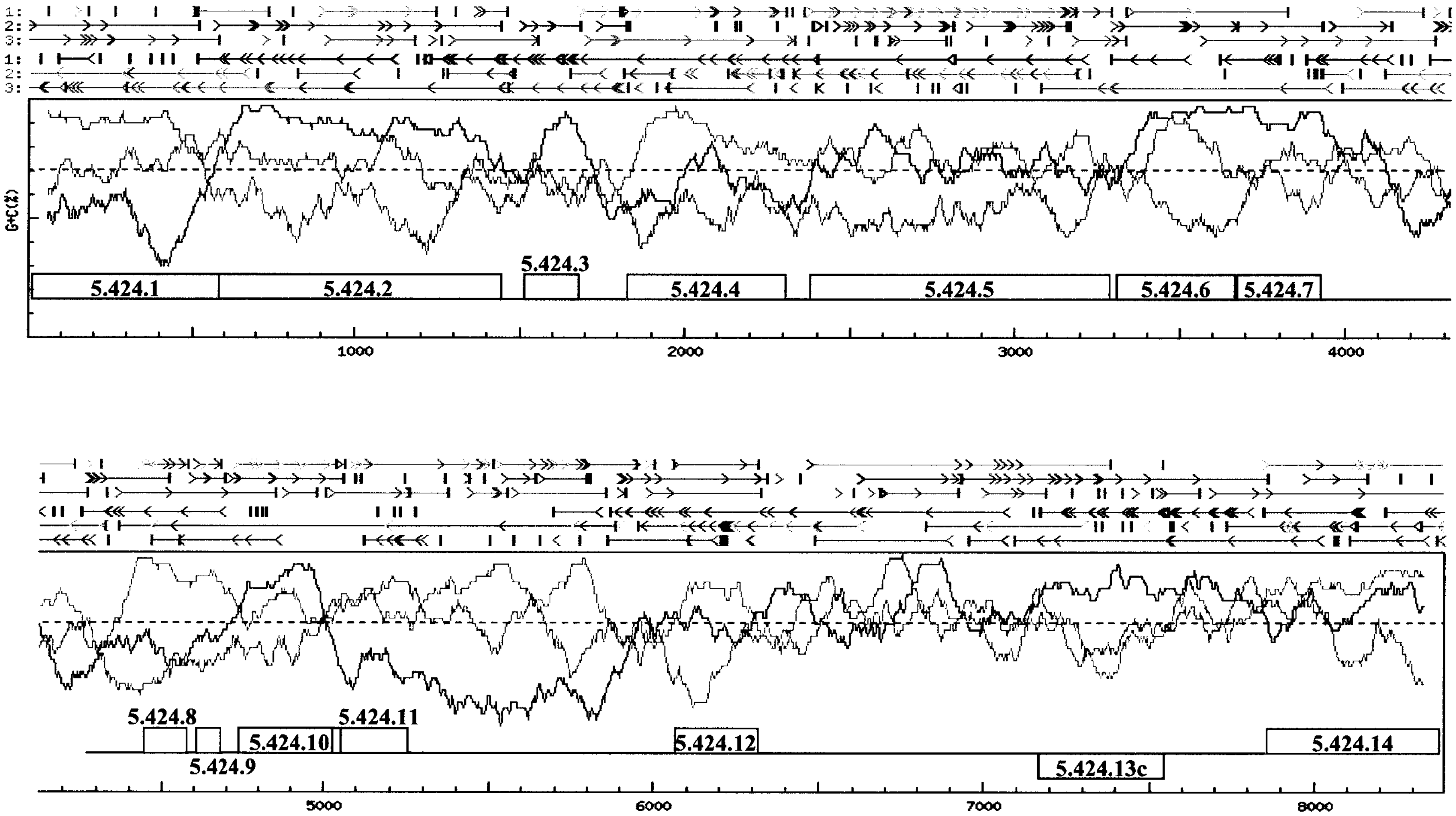


Contig 4.H9

FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig H9 5599 bp; 70.9% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 4 07:03:35 2003

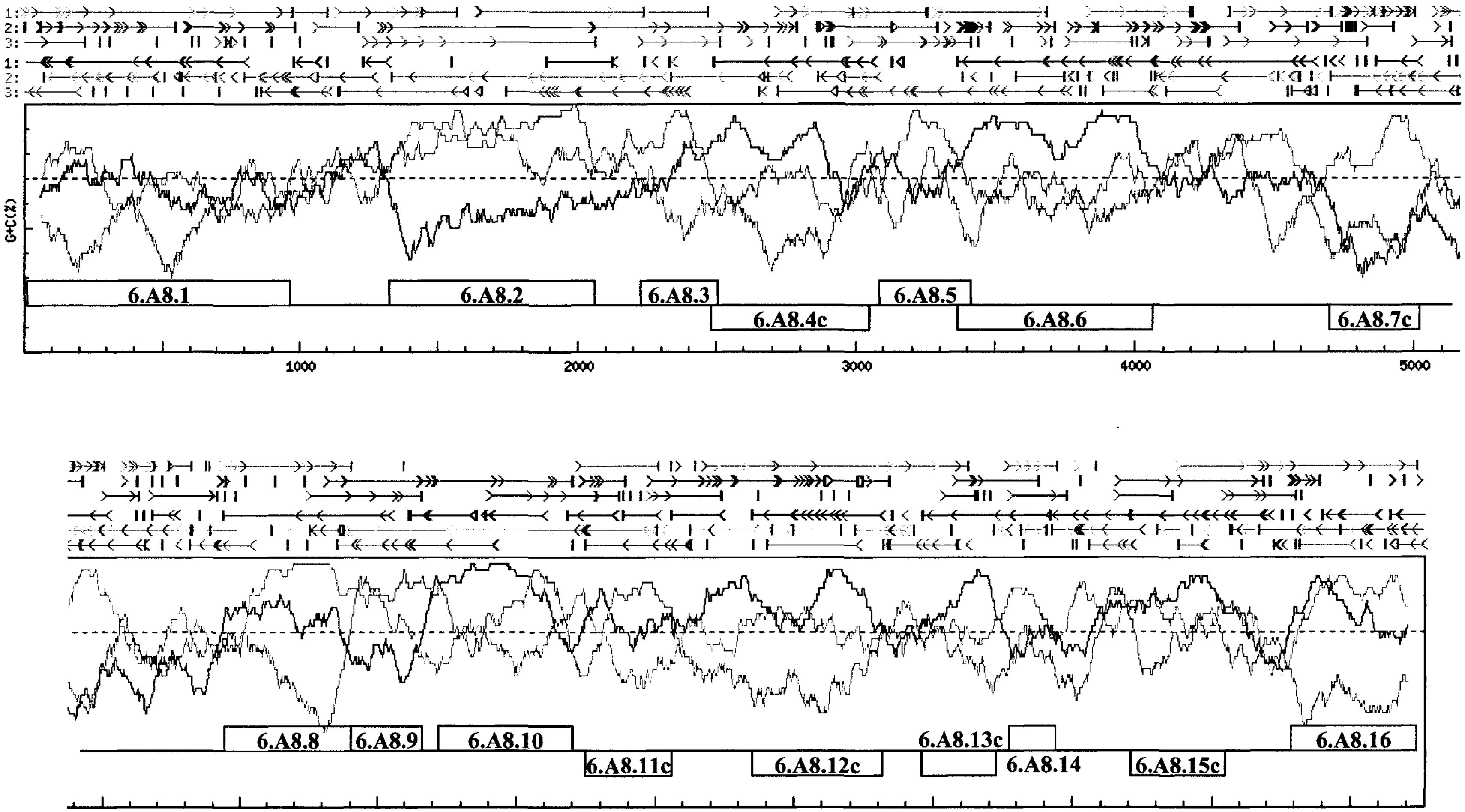
Contig 5.424

```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 424 8391 bp; 70.4% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 4 07:27:09 2003
```



Contig 6.A8

FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig A8 9765 bp; 69.7% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]; ATG GTG TTG
Minimum ORF: 20, Date: Jun 4 08:40:37 2003

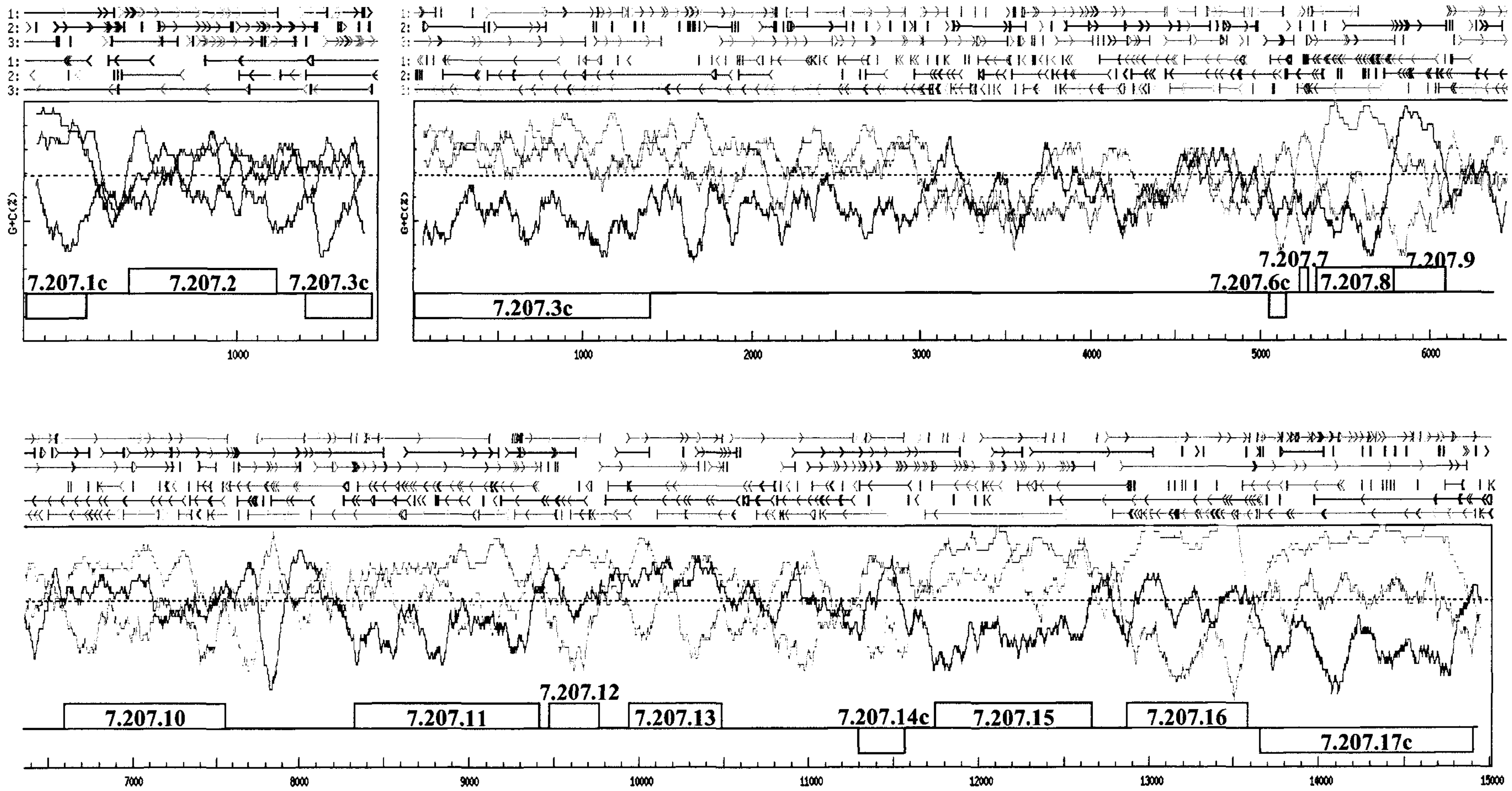


Contig 7.207-I

FramePlot 3,0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 207-I 1656 bp; 68,8% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 4 13:05:52 2003

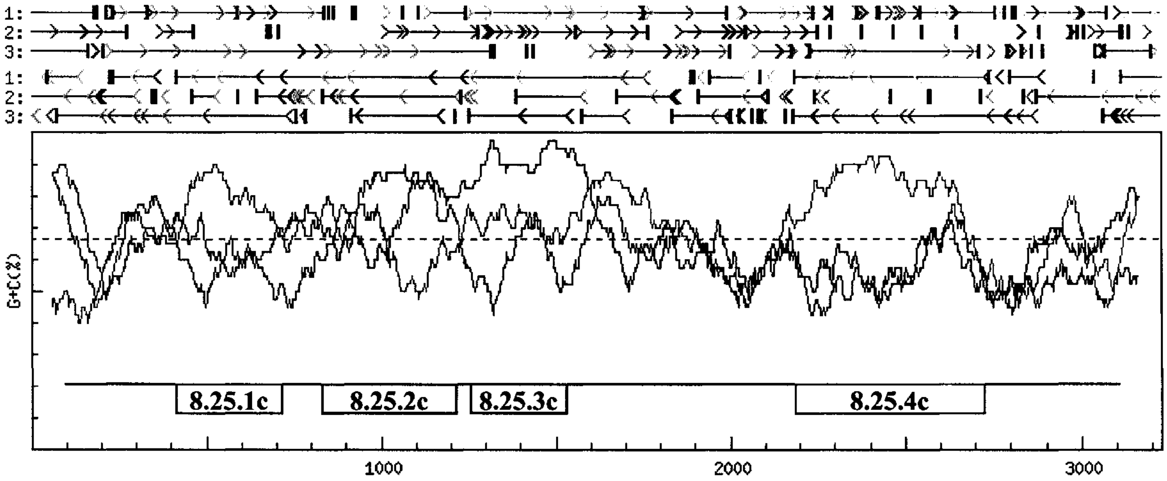
Contig 7.207-II

FramePlot 3,0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 207-II 15013 bp; 68,7% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 10 , Date: Jun 4 12:55:39 2003



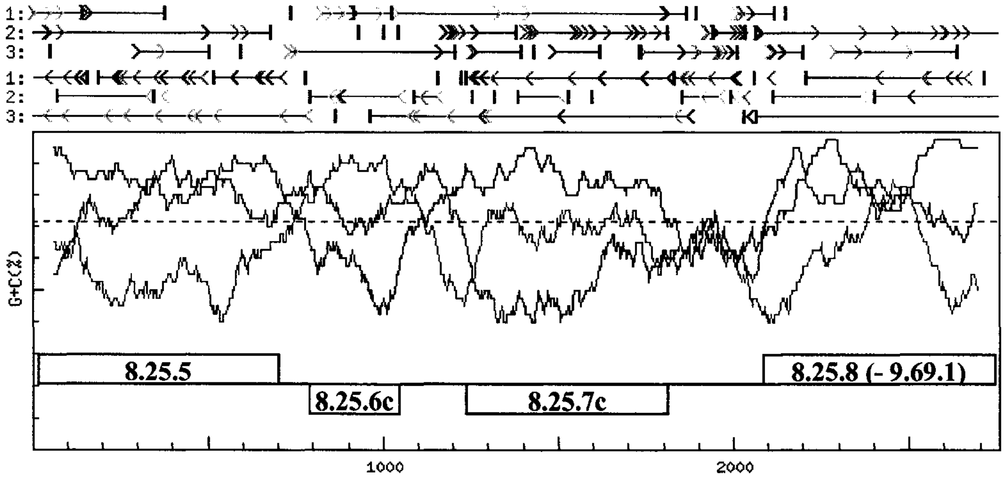
Contig 8.25-I

```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 25-I 3224 bp; 66.5% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 5 01:32:15 2003
```

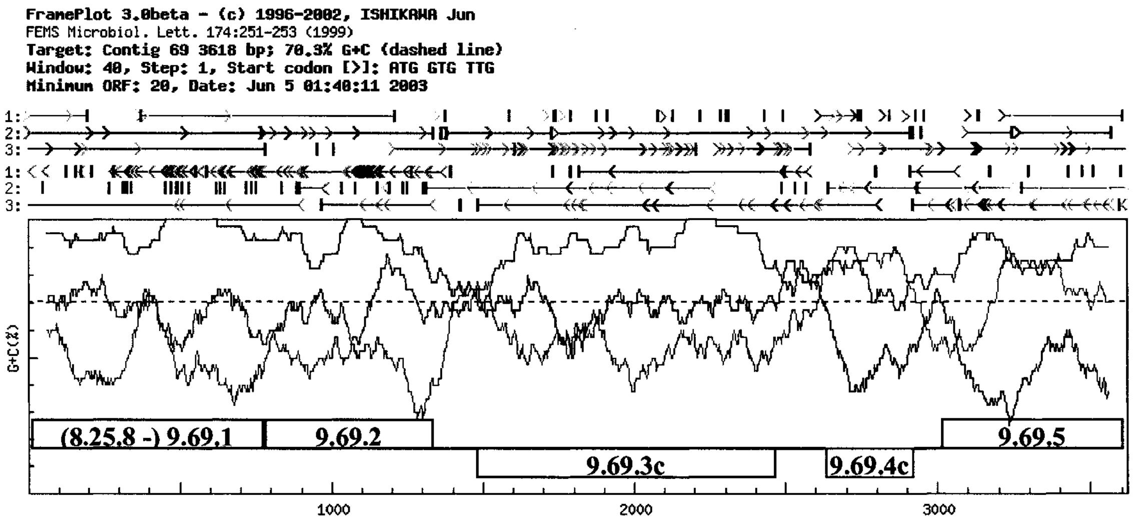


Contig 8.25-II

```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 25-II 2754 bp; 71.5% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 5 01:34:26 2003
```

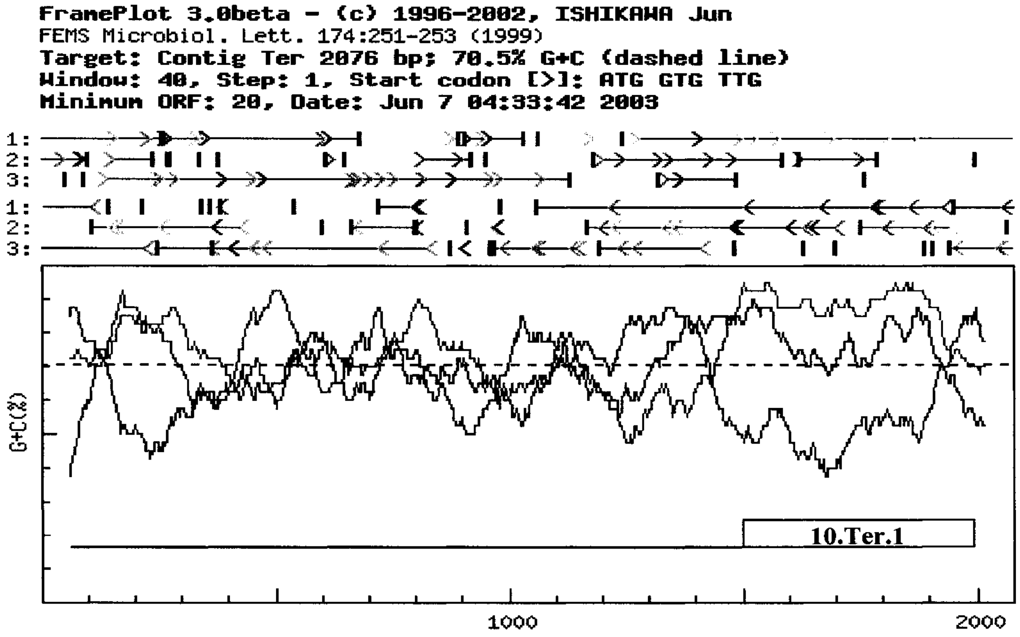


Contig 9.69



```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig 69 3618 bp; 70.3% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 5 01:40:11 2003
```

Contig 10.Ter



```
FramePlot 3.0beta - (c) 1996-2002, ISHIKAWA Jun
FEMS Microbiol. Lett. 174:251-253 (1999)
Target: Contig Ter 2076 bp; 70.5% G+C (dashed line)
Window: 40, Step: 1, Start codon [>]: ATG GTG TTG
Minimum ORF: 20, Date: Jun 7 04:33:42 2003
```

sequence (usually a variation on GGAGG) located 5 to 12 bp upstream of the translation start codon of a protein-coding gene. Although some examples have been found in *Streptomyces* species of protein-coding ORFs that lack an RBS, any ORF with a start codon that has a potential RBS upstream is more likely to be a protein-coding ORF.

3.5.2 Overview of the Hypothetical Proteins

The predicted protein-coding ORFs have been annotated by protein sequence similarity comparison and protein family classification. In total 97 ORFs detected in pSCL2 were annotated, of which 15 ORFs encode hypothetical proteins, which have similarity to other hypothetical proteins in GenBank but have unknown functions. Approximately half of the predicted ORFs (49 of them) encode unknown proteins, which have no similarity to any proteins in GenBank. A graphical representation of the ORFs in each contig is depicted in Fig 3.5-1, corresponding to their GC distribution in the frame plot. The predicted features of each ORF and its putative function by database analysis are listed in Table 3.5-1. In the next section, the characteristics of the putative proteins that have essential or important functions are discussed in detail.

3.6 CHARACTERIZATION OF THE PUTATIVE PROTEINS

3.6.1 Replication Proteins and Replication Origin

The capacity to replicate autonomously is an important feature for a plasmid. This section presents the computational identification, characterization, and experimental analysis of the function of two putative replication proteins and the replication origin in the linear plasmid pSCL2.

3.6.1.1 Identification of replication proteins of pSCL2

Analysis of a region of the assembled sequence from near the centre of pSCL2 revealed the presence of two complete ORFs (Fig. 3.6.1-1A), ORF 7.207.8 and ORF 7.207.9, that showed 86% and 74% identity to the genes encoding RepL1 and RepL2

Table 3.5.1. Features of the ORFs found on the linear plasmid pSCL2.

ORF (size [a])	Position (bp) (Start..Stop codons)	Most probable homolog (size [b])	Source of the homologs	% Identity [c]	Conserved Domains
1.68.1 (240)	3(?[d])..722(TGA)	Conserved hypothetical protein (238)	*S. avermitilis*	27 (33/118)	
1.68.2 (296) [type Ib *parA*]	722(ATG)..1609(TGA)	Putative cell division protein ParA (309)	*Corynebacterium glutamicum* ATCC 13032	26 (54/204)	
1.68.2a (100) [type Ib *parB*]	1658(ATG)..1960(TGA)	- [e]	- [e]	- [e]	
1.68.3 (92)	2017(ATG)..2292(TGA)	-	-	-	
1.68.4 (102)	2648(GTG)..2953(TGA)	Putative ribonuclease H (235)	*S. coelicolor* A3(2)	39 (22/56)	
1.68.5c[f] (134)	3474(GTG)..3073(TAG)	-	-	-	
1.68.6 (466)	4187(GTG)..5584(TGA)	-	-	-	
1.68.7 (193)	5994(GTG)..6575(TGA)	Hypothetical protein; SCP1.194 (185)	*S. coelicolor* A3(2)	51 (92/178)	
1.68.8c (102)	6948(ATG)..6640(TGA)	Putative secreted protein (132)	*S. coelicolor* A3(2)	31 (42/132)	
2.26.1 (303)	2(?)..910(TAG)	Helicase; SAP1_95 (885)	*S. avermitilis*	75 (230/303)	Helicase associated domain
2.26.2 (136)	1145(GTG)..1552(TGA)	Conserved hypothetical protein; SAP1_96 (125)	*S. avermitilis*	43 (54/123)	
2.26.3c (134)	2065(ATG)..1664(TGA)	-	-	-	
2.26.4 (271)	2377(GTG)..3189(TAG)	-	-	-	
2.26.5 (100)	3267(GTG)..3566(?[d])	-	-	-	
2.26.6 (774)	1(?)..2322(TAG)	Helicase (854)	*S. coelicolor* A3(2)	35 (50/142)	HELICc
2.26.7c (152)	2841(ATG)..2386(TGA)	Conserved hypothetical protein; pSV2.21c (218)	*S. violaceoruber*	46 (63/137)	
2.26.8 (527)	4342(GTG)..5925(TAG)	Ala-rich protein; SAP1_44 (734)	*S. avermitilis*	26 (80/298)	
		Large Ala/Glu-rich protein (1326)	*S. coelicolor*	26 (75/287)	
2.26.10 (63)	6106(GTG)..6294(TAG)	-	-	-	
2.26.11 (71)	6853(GTG)..7065(TGA)	Hypothetical protein; SAP1_80 (145)	*S. avermitilis*	42 (24/56)	
		Hypothetical protein; pSV2.24 (154)	*S. violaceoruber*	38 (22/57)	

2.26.12c (113)	8114(GTG)..7776(TGA)	Hypothetical protein; SAP1_76 (115)	*S. avermitilis*	71 (81/114)	
		Hypothetical protein; pSV2.26c (115)	*S. violaceoruber*	68 (78/114)	
2.26.13c (606)	9931(GTG)..8114(TGA)	Putative traA like protein; SAP1_75 (608)	*S. avermitilis*	79 (488/613)	
		Putative transfer protein; pSV2.27c (269) + hypothetical protein; pSV2.28c (333)	*S. violaceoruber*	84 (229/270), 75 (256/338)	
2.26.14 (80)	10074(GTG)..10313(TGA)	Hypothetical protein; pSV2.29 (72)	*S.violaceoruber*	38 (26/67)	
3.73.1 (56)	9(ATG)..176(TAG)	-	-	-	
3.73.2 (287)	189(ATG)..1052(TGA)	Tra3 (339)	*S. bambergiensis*	45 (105/232)	DUF721
3.73.4 (137)	1366(GTG)..1776(TGA)	pBL1 hypothetical protein 7 (109)	*S. bambergiensis*	33 (30/89)	
3.73.5 (1009)	2063(GTG)..5089(TGA)	Putative ATP/GTP binding protein (966)	*S. coelicolor*	29 (267/898)	TPR, NB-ARC
3.73.6 (279)	5332(GTG)..6168(TAG)	IS*468* Transposase (279)	*S. coelicolor*	94.5 (264/279)	COG3293 Transposase DDE
3.73.7 (132)	6495(GTG)..6890(TGA)	Putative acetyltransferase (179)	*S. coelicolor*	50 (18/36)	
3.73.8c (317)	7938(ATG)..6988(TGA)	Hypothetical protein (315)	*S. avermitilis*	36 (107/296)	
3.73.9 (268)	8230(GTG)..9033(TGA)	-	-	-	
4.H9.1 (70)	12(ATG)..221(TGA)	-	-	-	
4.H9.2 (100)	450(ATG)..749(TGA)	-	-	-	
4.H9.3 (85)	749(GTG)..1003(TGA)	-	-	-	
4.H9.4 (244)	1378(GTG)..2109(TGA)	Putative GntR-family regulatory protein (252)	*S. coelicolor* A3(2)	29 (70/240)	HTH_GNTR, PhnF
		Xylanase regulatory protein (252)	*S. lividans*	28 (69/240)	
		pSCL1 ORF-L (248)	*S. clavuligerus*	21 (47/221)	
4.H9.5c (39)	2667(GTG)..2551(TAG)	-	-	-	
4.H9.6 (158)	3209(ATG)..3682(TGA)	-	-	-	
4.H9.7 (100)	3934(ATG)..4233(TGA)	-	-	-	
4.H9.8 (86)	4233(GTG)..4490(TGA)	-	-	-	
4.H9.9c (309)	5597(?)..4671(TAG)	Putative ATP/GTP binding protein (966)	*S. coelicolor* A3(2)	41 (125/298)	
		Serine/threonine-protein kinase (754)	*S. avermitilis*	41 (122/297)	
5.424.1 (193)	3(?)..581(TGA)	Putative RNA polymerase sigma factor; SAP1_87 (301)	*S. avermitilis*	42 (81/191)	RpoE, sigma24 homolog

		Hypothetical protein; pSV2.08c (529)	*S. violaceoruber*	39 (75/188)	
5.424.2 (287)	581(GTG)..1441(TGA)	Conserved hypothetical protein; SAP1_88 (291)	*S. avermitilis*	37 (113/299)	
		Hypothetical protein; pSV2.08c (529)	*S. violaceoruber*	38 (99/259)	
5.424.3 (55)	1520(ATG)..1684(TGA)	-	-	-	
5.424.4 (160)	1825(ATG)..2304(TAG)	Hypothetical protein (164)	*Magnetospirillum magnetotacticum*	48 (79/164)	Acetyltransferase (GNAT) family
		Putative MarR-family transcriptional regulator (310)	*S. avermitilis* (SAV5311)	29 (32/109)	
5.424.5 (266)	2386(GTG)..3183(TAA)	Putative phosphatase (274)	*S. coelicolor*	74 (193/259)	SuhB, inositol_P
5.424.6 (121)	3311(TTG)..3673(TAG)	-	-	-	
5.424.7 (85)	3677(ATG)..3931(TGA)	Putative two-component regulator (206)	*S. viridochromogenes*	31 (25/80)	HTH_LUXR
5.424.8 (46)	4444(GTG)..4584(TAG)	-	-	-	
5.424.9 (23)	4615(TTG)..4686(TAG)	-	-	-	
5.424.10(104)	4748(ATG)..5062(TGA)	-	-	-	
5.424.11 (74)	5037(TTG)..5261(TAG)	-	-	-	
5.424.12 (81)	6076(ATG)..6321(TGA)	-	-	-	
5.424.13c (123)	7539(ATG)..7168(TGA)	-	-	-	
5.424.14(176)	7861(GTG)..8391(?)	-	-	-	
6.A8.1 (325)	1(?)..978(TAG)	Conserved hypothetical protein; SAP1_29 (716)	*S. avermitilis*	34 (104/304)	
6.A8.2 (243)	1332(GTG)..2063(TGA)	-	-	-	
6.A8.3 (93)	2235(ATG)..2516(TGA)	-	-	-	
6.A8.4c (151)	2940(TTG)..2485(TAA)	-	-	-	
6.A8.5 (104)	3099(GTG)..3413(TGA)	Putative Tra3 homolog; SAP1_34 (161)	*S. avermitilis*	55 (25/45)	
		pBL1 Tra3 (339)	*S. bambergiensis*	54 (18/33)	
6.A8.6c (232)	4053(ATG)..3355(TGA)	Putative secreted protein (219)	*S. coelicolor* A3(2)	59 (131/221)	
		Putative secreted protein; pSV2.11 (222)	*S. violaceoruber*	51 (114/222)	
6.A8.7c (104)	5012(GTG)..4698(TGA)	-	-	-	
6.A8.8 (128)	5524(ATG)..5910(TGA)	-	-	-	
6.A8.9 (85)	5907(ATG)..6164(TGA)	Conserved hypothetical protein (96)	*Vibrio cholerae*	25 (22/85)	
6.A8.10 (159)	6224(GTG)..6703(TAG)	Putative Tra3 protein; SAP1_32 (110)	*S. avermitilis*	61 (39/63)	
		pBL1 Tra3 (339)	*S. bambergiensis*	47 (42/88)	

6.A8.11c (104)	7054(ATG)..6740(TGA)	Putative outer membrane protein TprK (496)	*Treponema pallidum subsp. pallidum*	34 (17/50)	
6.A8.12c (154)	7806(ATG)..7342(TAG)	-	-	-	
6.A8.13c (86)	8217(TTG)..7957(TGA)	-	-	-	
6.A8.14 (58)	8272(TTG)..8448(TGA)	-	-	-	
6.A8.15c (99)	9000(GTG)..8701(TGA)	-	-	-	
6.A8.16 (145)	9304(ATG)..9741(TGA)	Conserved hypothetical protein; pSV2.105c (146)	*S. violaceoruber*	91 (134/146)	
		Hypothetical protein (152)	*S. coelicolor* A3(2)	89 (128/143)	
7.207.1c (65)	201(GTG)..4(?)	-	-	-	
7.207.2 (236)	487(ATG)..1197(TGA)	-	-		
7.207.3c (467..102)	(207-II)-1406(GTG)..3 (207-I)-1628..1323(TGA)	Putative SNF2/RAD54 family helicase (962)	*S. avermitilis* MA-4680	31 (137/435) 41 (41/98)	HepA, Superfamily II DNA/RNA helicases
7.207.6c (34)	5158(GTG)..5054(TGA)	-	-	-	
7.207.7 (13)	5242(GTG)..5283(TAG)	-	-	-	
7.207.8 (150) [*repC1*]	5340(TTG)..5792(TGA)	pSLA2-L replicative protein RepL1 (150)	*S. rochei*	86 (130/150)	
7.207.9 (102) [*repC2*]	5789(GTG)..6097(TGA)	pSLA2-L replicative protein RepL2 (104)	*S. rochei*	74 (76/104)	
7.207.10 (321)	6598(GTG)..7563(TGA)	Hypothetical protein SC2E1.30 (183)	*S. coelicolor* A3(2)	30 (54/180)	
7.207.11 (363)	8334(ATG)..9425(TAG)	Hypothetical protein (413)	*S coelicolor* A3(2)	35 (46/129)	
7.207.12 (97)	9481(GTG)..9774(TAA)	-	-	-	
7.207.13 (181)	9946(ATG)..10491(TGA)	Putative Tra3 protein; SAP1_32 (110)	*S. avermitilis*	61 (34/55)	
		pBL1 hypothetical protein 3 (339)	*S. bambergiensis*	65 (28/43)	
7.207.14c (97)	11587(ATG)..11294(TGA)	Putative muconolactone delta-isomerase (217)	*S. avermitilis* MA-4680	81 (77/94)	Muconolactone delta-isomerase
7.207.15 (311)	11745(GTG)..12680(TGA)	Putative lysR-family transcriptional regulator (304)	*S. avermitilis* MA-4680	75 (230/306)	LysR
7.207.16 (236)	12883(ATG)..13593(TGA)	Putative oxidoreductase (237)	*S. avermitilis* MA-4680	84 (195/232)	FabG, Dehydrogenases
7.207.17c (400)	14846(TTG)..13644(TGA)	IS*116* transposase (399)	*S. clavuligerus*	99.8 (398/400)	Transposase_20, Transposase_9

8.25.1c (99)	707(ATG)..408(TAA)	-	-	-	
8.25.2c (127)	1210(GTG)..827(TGA)	-	-	-	
8.25.3c (93)	1530(ATG)..1249(TGA)	-	-	-	
8.25.4c (179)	2720(ATG)..2181(TGA)	-	-	-	
8.25.5 (226)	2(?)..682(TAG)	-	-	-	
8.25.6c (84)	1043GTG)..789(TGA)	-	-	-	
8.25.7c (192)	1809(ATG)..1231(TGA)	Putative methyltransferase (548)	*Deinococcus radiodurans*	43 (58/132)	COG4122, Predicted O-methyltransferase
		Putative transferase (203)	*S. coelicolor* A3(2)	33 (49/148)	
8.25.8 (220) [*tapCL1*]	2090(GTG)..2752(?)	pSLA2-M telomere-binding protein TapR2 (734)	S. rochei	68 (152/223)	Helix-turn-helix motif (HTH_3)
9.69.1 (256) [*tapCL1*]	2(?)..772(TAG)	Putative transcriptional regulator; pSV2.81 (656)	*S. violaceoruber*	89 (229/256)	
		TpgARg2; SAP1_22 (772)	*S. avermitilis*	86 (221/255)	
		Telomere-binding protein TapC (739)	*S. coelicolor*	85 (220/256)	
9.69.2 (185) [*tpgCL1*]	785(ATG)..1342(TAG)	Putative terminal protein Tpga2; SAP1_21 (185)	*S. avermitilis*	75 (138/184)	HTH
9.69.3c (323)	2452(ATG)..1481(TGA)	-	-	-	
9.69.4c (94)	2921(ATG)..2637(TAG)	-	-	-	
9.69.5 (196)	3027(ATG)..3617(?)	-	-	-	
10.Ter.1 (-)	1417(ATG)..?(?)	-	-	-	

a. The numbers in parentheses describe the sizes of the ORF products (aa).

b. The numbers in parentheses describe the sizes of the homologs (aa).

c. The fractions in parentheses represent the number of identical amino acid over their alignment region.

d. "?" represents that the coding region is lacking the start or stop codon.

e. "-" represents no significant similarity.

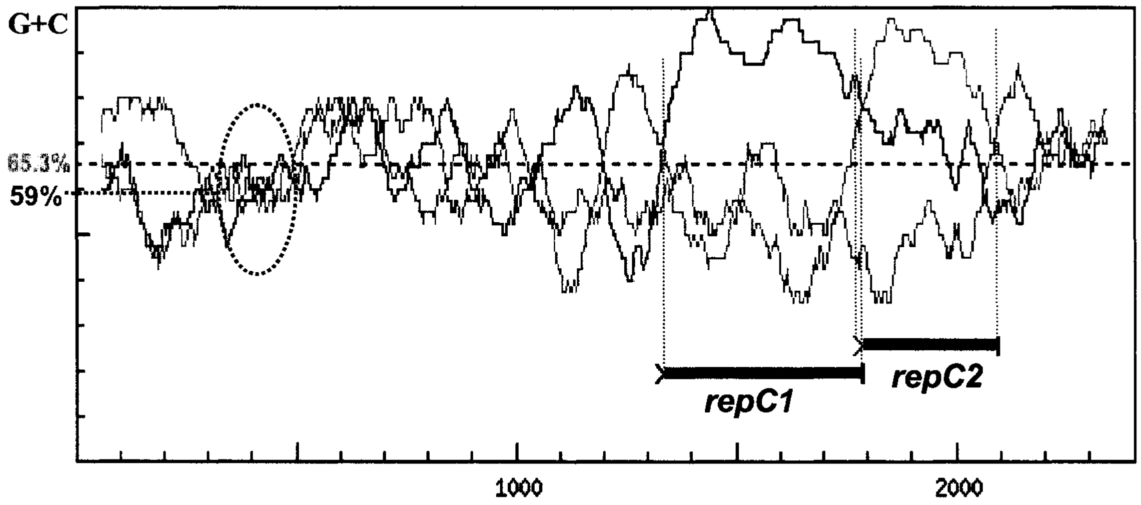
f. c, transcription in the other direction.

Fig. 3.6.1-1. Identification of the replication genes in pSCL2 and pSLA2-L by the FramePlot program.

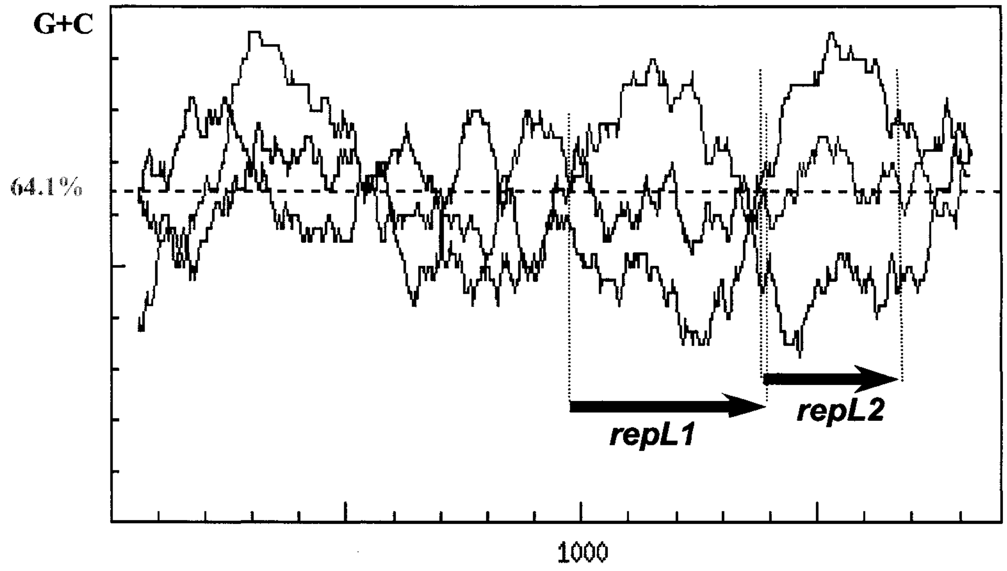
(A) The replication genes *repC1* and *repC2* were predicted in pSCL2. The *repC1* gene is represented by the green bar and the green line in the plot with an average 90.7% GC content in the 3rd-letter position. The *RepC2* gene is represented by the red bar and the red line in the plot with an average 89.3% GC content in the 3rd-letter position. The average GC content of the replication region in pSCL2 is 65.3% and the GC content of the circled region (Box 3 in Fig. 3.6.1-3) is about 59%.

(B) The replication genes *repL1* and *repL2* were predicted in pSLA2-L. The *repL1* gene is represented by the red bar and the red line in the plot with an average 75.5% GC content in the 3rd-letter position. The *RepL2* gene is represented by the blue bar and the blue line in the plot with an average 85.7% GC content in the 3rd-letter position. The average GC content of the replication region in pSLA2-L is 64.1%.

(A) pSCL2



(B) pSLA2-L



respectively (Fig. 3.6.1-1B), of pSLA2-L, a linear plasmid in *Streptomyces rochei* (Hiratsu *et al.*, 2000). No significant similarities were found to any other sequences in the GenBank database. These two ORFs were located adjacent to each other and both were oriented in the same direction, just as are the genes encoding RepL1 and RepL2 in pSLA2-L. RepL1 (150 amino acids) and RepL2 (104 amino acids) have been shown to be replication proteins for pSLA2-L (206 kb). Based on the results of this homology search, we predicted that the two ORFs found in pSCL2 would also encode replication proteins. These ORFs were designated as the *repC1* and *repC2* genes, which encode the RepC1 (150 amino acids) and RepC2 (102 amino acids) proteins, respectively.

Comparing the predicted properties of the RepC and RepL proteins (Table 3.6.1-1), The *repC* and *repL* genes are not only similar in DNA sequence and length, but also share all of the features required for their translation, including start and stop codons, potential ribosome binding sites (RBS), and an acceptable distance from the RBS to the start codon. The predicted physical characteristics of the putative RepC and RepL proteins were compared, which included hydrophilicity, beta-sheet, alpha-helix and flexible regions. Each of these parameters was very similar for the two pairs of proteins suggesting that they share very similar secondary structures as well (Fig. 3.6.1-2). Their tertiary structures were predicted using an online program, 3D-PSSM. Both RepC1 and RepL1 are alpha proteins. They have similarity to the proteins in the superfamily containing a “winged helix” DNA-binding domain (Brennan, 1993), consistent with their proposed involvement in the initiation of replication. RepC2 and RepL2 are also alpha proteins but more distantly related to DNA-binding proteins. Such marked similarities give further credence to the proposal that the RepC proteins could function as replication proteins in the same way that the RepL proteins do.

3.6.1.2 Putative replication origin for pSCL2

Previous studies of other *Streptomyces* linear plasmids and chromosomes have shown that their origins of replication are normally located upstream of replication proteins, and contain a series of direct and/or inverted repeats and low GC content sequences (Redenbach *et al.*, 1999; Hiratsu *et al.*, 2000). Based on such common

Table 3.6.1-1. Comparison of translation features and other predicted properties of the RepC proteins of pSCL2 (*S. clavuligerus*) and the RepL proteins of pSLA2-L (*S. rochei*).

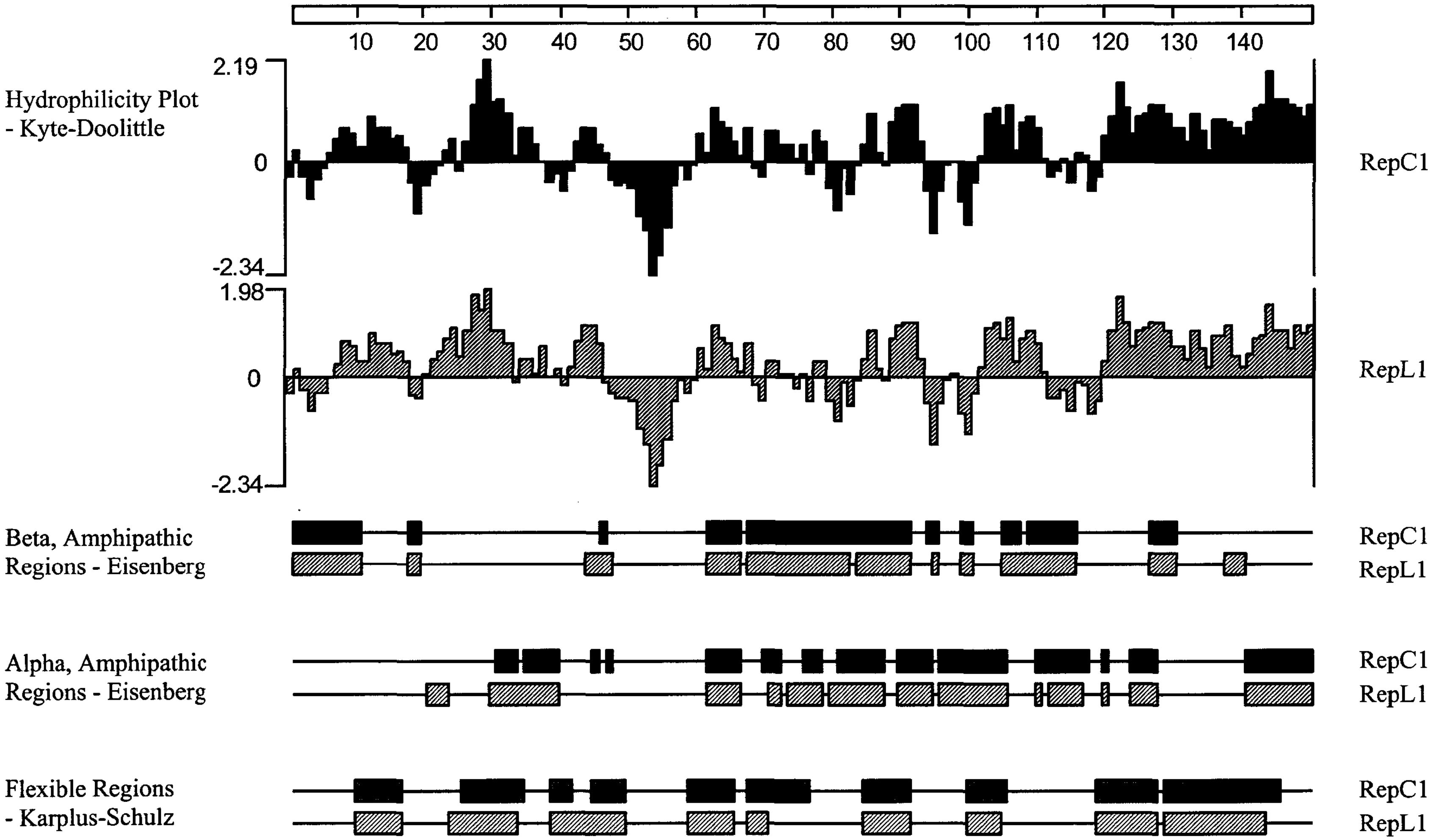
	RepC1	RepL1	RepC2	RepL2
Potential translation initiation codon	TTG	TTG	GTG	GTG
Potential ribosomal-binding site (RBS)	GGAGG	GGAGG	GGAGG	GGAGG
Distance from the RBS to the initiation codon	5 bp	5 bp	4 bp	4 bp
Termination codon	UGA	UGA	UGA	UGA
Length (aa)	150	150	102	104
Molecular mass (kDa)	16.64	16.67	11.48	11.66
Isoelectric point	9.54	9.53	10.14	9.28
Charge at pH 7	+ 6.36	+ 5.39	+ 5.42	+ 2.42
Homology	86%		74%	

Fig. 3.6.1-2. Comparison of the predicted secondary structures of the putative RepC and RepL proteins.

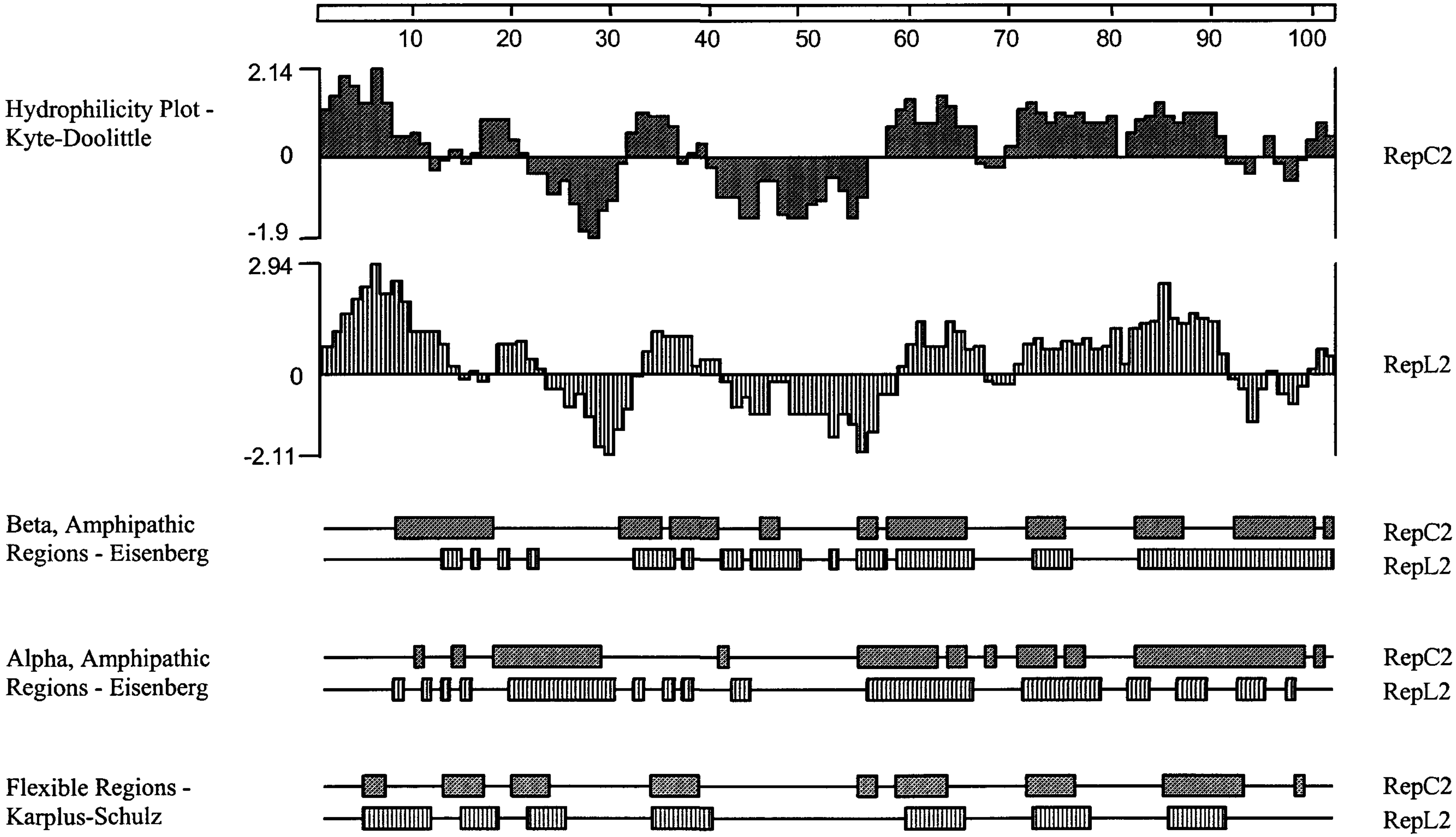
A. Comparison of the RepC1 and RepL1 proteins. The upper plots relate to RepC1, and the bottom plots relate to RepL1.

B. Comparison of the RepC2 and RepL2 proteins. The upper plots relate to RepC2, and the bottom plots relate to RepL2.

A. Comparison of RepC1 and RepL1



B. Comparison of RepC2 and RepL2



features, three regions upstream of *repC1* were identified which could function as the replication origin.

The first hypothetical replication origin (1050-850 bp upstream of *repC1*; Box 1 in Fig. 3.6.1-3) has a relatively low GC content, which drops to 59% from the average of 65.3% in the replication region, including the replication genes (Fig. 3.6.1-1A). A low GC content has been suggested to facilitate melting of the double-stranded DNA in this region and thereby help to initiate replication. No long direct or inverted repeats were discovered in this region. The sequence in the second hypothetical replication region (720-615 bp upstream of *repC1*; Box 2 in Fig. 3.6.1-3) shows 78% identity to the putative pSLA2-L replication origin. Both pSCL2 and pSLA2-L contain two pairs of palindromes in this region but do not show the low GC content typical of other replication origins (Hiratsu *et al.*, 2000). The third possible region (224-168 bp upstream of *repC1*; Box 3 in Fig. 3.6.1-3) contains the longest palindrome close to *repC1*, with a perfect 26 bp stem and 5-base loop. Such inverted repeats could serve as recognition sites for replication proteins involved in initiation of replication.

3.1.6.3 Functional analysis of replication proteins and replication origin of pSCL2

In order to gain further evidence for the involvement of RepC1 and RepC2 and the surrounding DNA sequence in the replication of pSCL2, a functional analysis of the replication origin region was carried out. The sequencing library of pSCL2 (constructed in pCR®4Blunt-TOPO®) was searched for suitable clones that cover the predicted replication origin region of pSCL2. Two clones, pSG207 (2485 bp upstream of *repC1* to 198 bp downstream of *repC2*) and pSGH4 (1030 bp upstream of *repC1* to 693 bp downstream of *repC2*), both covering the complete *repC1* and *repC2* gene sequences and upstream hypothetical replication origin regions, were selected for this purpose (Fig. 3.6.1-4). Since the pCR®4Blunt-TOPO® vector does not carry a *Streptomyces* origin of replication, the plasmids would only be expected to survive and confer neomycin resistance if the inserts contained functional origins of replication. pSG230, included as a negative control, contained a pSCL2 DNA fragment from outside of the putative replication origin cloned into the pCR®4Blunt-TOPO® vector. A positive control, pST1,

Fig. 3.6.1-3. Distinctive features in the replication origin region of pSCL2.

A: Genetic organization of the replication origin region of pSCL2. The putative genes, *repC1* and *repC2*, are indicated by black arrows. The open boxes 1, 2 and 3 represent three hypothetical replication origin elements.

B: The sequence in box 2 of pSCL2 and its alignment with the putative replication origin of pSLA2-L. Palindromes are shown as pairs of arrows with solid or broken lines. The bold and italic letters show the sequence of the putative replication origin of pSLA2-L.

C: The sequence within box 3 of pSCL2. A perfect palindrome with a 26-bp stem and 5-base loop is indicated by arrows above the sequence. This palindrome is located 224-168 bp upstream of the *repC1* gene.

The sequence data for the replication region of pSCL2 is included in Contig207-II that has been submitted to the NCBI GenBank database and assigned Accession Number AY392418.

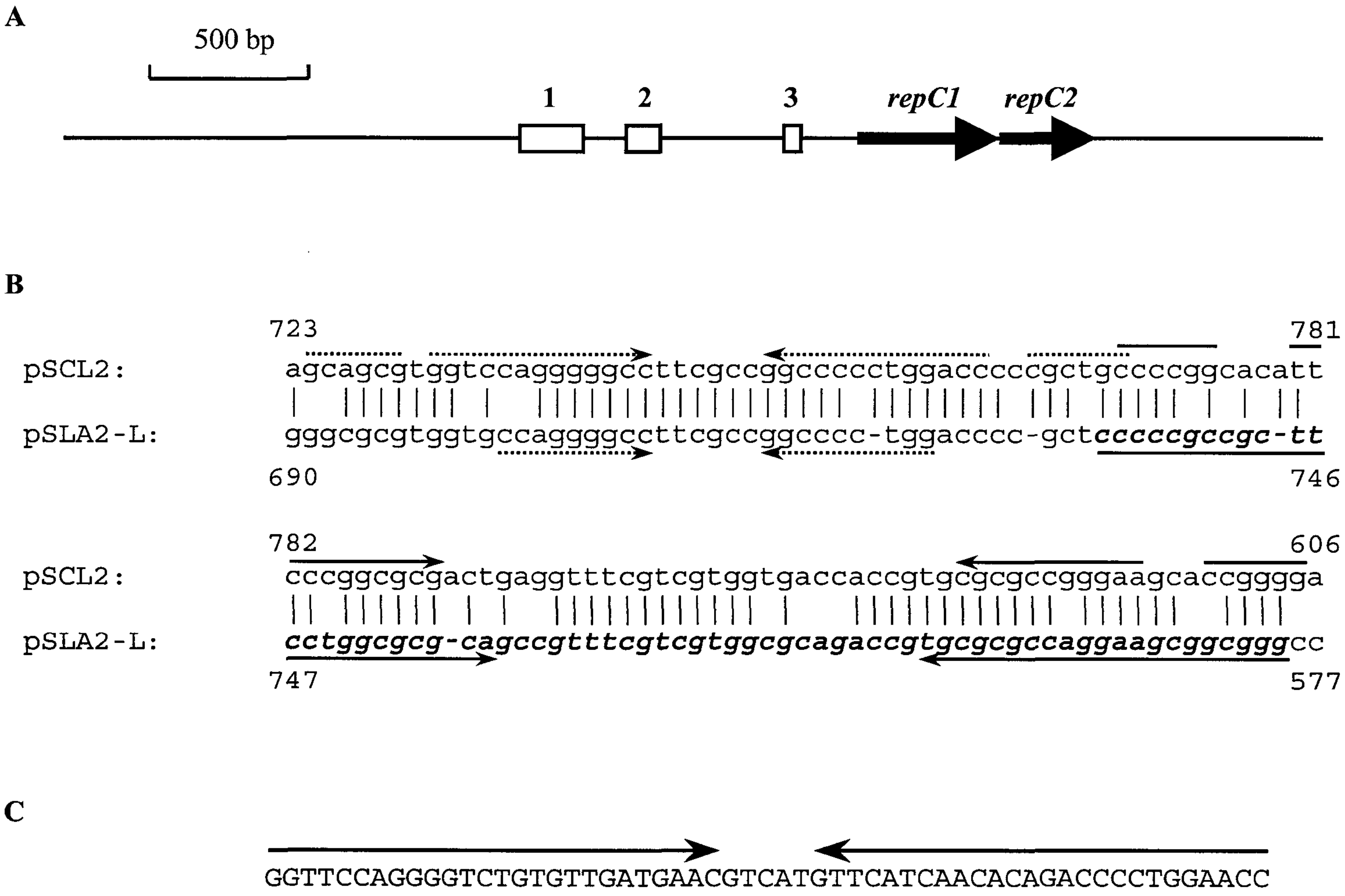
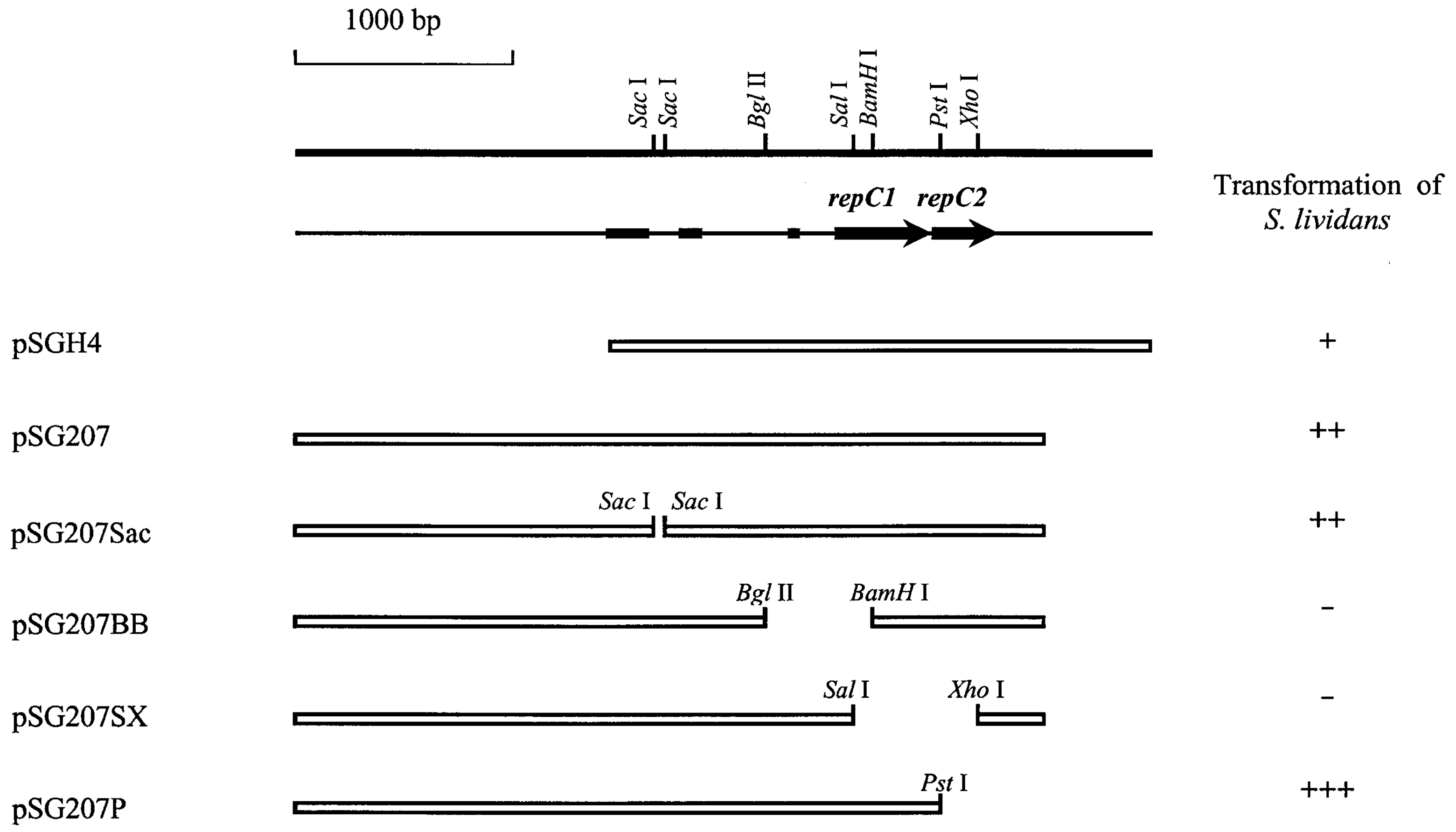

A
500 bp
1
2
3
repC1
repC2
B
723
781
pSCL2:
agcagcgtggtccagggggccttcgccggccccctggaccccgctgccccggcacatt
pSLA2-L:
gggcgcgtggtgccaggggccttcgccggcccc-tggacccc-gctcccccgccgc-tt
690
746
782
606
pSCL2:
cccggcgcgactgaggtttcgtcgtggtgaccaccgtgcgcgccgggaagcaccgggga
pSLA2-L:
cctggcgcg-cagccgtttcgtcgtggcgcagaccgtgcgcgccaggaagcggcgggcc
747
577
C
GGTTCCAGGGGTCTGTGTTGATGAACGTCATGTTCATCAACACAGACCCCTGGAACC

Fig. 3.6.1-4. Location of pSCL2 replication origin sequences cloned in pSGH4, pSG207 and derivatives of pSG207.

The locations of DNA sequences cloned into pSGH4, pSG207, and four deletion derivatives of pSG207, are indicated with respect to the putative origin of replication region. The ability of the various plasmids to replicate in *S. lividans* is also indicated. The restriction sites used in the deletion analysis are shown at the top. The putative replication genes and the three hypothetical origin elements are indicated by black arrows and red boxes, respectively. The "+" and "-" symbols indicate the ability of each plasmid to replicate in *S. lividans*. The number of "+" represents the transformation efficiency.


1000 bp
Sac I
Sac I
Bgl II
Sal I
BamH I
Pst I
Xho I
repC1
repC2
Transformation of S. lividans
pSGH4
+
pSG207
++
pSG207Sac
Sac I
Sac I
++
pSG207BB
Bgl II
BamH I
-
pSG207SX
Sal I
Xho I
-
pSG207P
Pst I
+++

was constructed by cloning the 3.1 kb fragment from the pHJL400 shuttle vector (Larson & Hershberger, 1986), which contains the replication origin of SCP2*, a circular plasmid from *S. coelicolor*, into pCR®-BluntII-TOPO® vector. The two plasmids carrying putative origins of replication, pSG207 and pSGH4 were introduced into *S. lividans* TK24 protoplasts by transformation along with the positive and negative control plasmids with selection for neomycin resistance. The DNA concentrations of all of the four plasmids were measured to ensure that same amount of DNA was transformed. As a result, pSG207, pSGH4 and pST1 transformants could be seen on plates supplemented with neomycin after 7 days incubation at 28°C, whereas no growth was seen on the plates receiving pSG230 transformants (Table 3.6.1-2).

Transformation efficiency is defined as the number of successfully transformed cells obtained over the amount of DNA transformed into competent cells. Since the transformed protoplasts were evenly spread on the R5 plates, each colony observed on plates after incubation should be able to represent one successfully transformed cell. The transformation efficiencies of pSG207, pSGH4 and pST1 were compared. More pSG207 colonies were observed on R5 plates than pSGH4 colonies, and both of them grew more slowly than pST1 transformants. Since the same amount of DNA was used to start transformation, this suggests that pSG207 has higher transformation efficiency than pSGH4; pSGH4 may lack certain sequences that are important to direct replication of pSCL2. The neomycin-resistant colonies were patched onto MYM plates, and then inoculated into TSBS liquid medium to grow mycelium. Free pSG207 and pSGH4 plasmids were isolated from 96-hour cultures and had the same sizes and restriction patterns as the original plasmids introduced into *S. lividans*. This indicate that the plasmids did not integrate into the *S. lividans* chromosome but were able to replicate autonomously.

To further confirm the function of the two replication genes in pSCL2, deletion analysis was carried out. As shown in Fig. 3.6.1-4, after deleting a *Bgl*II-*BamH*I (pSG207BB) or *Sal*I-*Xho*I (pSG207SX) fragment, neither plasmid could replicate in *S. lividans* TK24. This indicates that RepC1 is essential for the replication, and that *S. lividans*-derived replication proteins cannot replace RepC1 to support the replication of

Table 3.6.1-2. Transformation of plasmids carrying the putative pSCL2 replication origin region into *S. lividans*.

	R5 overlayed with Neo [c]	Patch on MYM with Neo [c]
pST1	++++ [b]	Confluent growth
pSG207	++	Scattered colonies
pSGH4	+	Scattered colonies
pSG230	- [b]	No colonies
BK[a]	-	No colonies

a. "BK" represents a background control, in which no plasmid was transformed.

b. The "+" and "-" symbols indicate the presence or absence of colonies on R5 plates. The number of "+" represents the transformation efficiency.

c. pST1, pSG207 and pSGH4 colonies, from R5 plates with neomycin (60 μg/ml), and pSG230 and BK colonies, from R5 plates without neomycin, were patched on MYM plates with 60 μg/ml neomycin.

pSCL2. In contrast, deletion of the RepC2 protein by excision of the *Pst*I fragment (pSG207P) did not prevent replication, but instead seemed to increase the transformation efficiency. This suggests that RepC2 is not essential for initiation of replication and may serve as a negative regulatory protein to maintain the low copy number of pSCL2. A comparison of colony growth, as an indication of transformation efficiency, of pSG207, pSG207P and the positive control, pST1, is shown in Fig. 3.6.1-5. The derived plasmid pSG207Sac was produced by deleting a small *Sac*I fragment that is located between the first and second hypothetical origin regions. The transformation efficiency of pSG207Sac did not differ significantly from that of the original plasmid, pSG207, suggesting that the spacing between these two regions is not important for initiation of replication.

3.6.2 Terminal Proteins

Linear plasmid replication is accomplished mainly by a bi-directional mechanism from the centrally located origin, except for the approximately 300-bp sequences at their telomeres, which are completed by synthesis initiated from the terminal proteins. Terminal proteins are unique to linear replicons and have been identified in pSCL2 as well. Frame analysis of the assembled sequence close to the right end of pSCL2 revealed the presence of two adjacent ORFs that have high similarities to the terminal proteins and the telomere-associated proteins found in many *Streptomyces* linear replicons.

ORF 9.69.2 encodes a 185 amino acid protein, with a theoretical molecular weight of 20.5 kDa and a pI of 10.07. The putative ribosome binding site (AGGAG) is located 9-bp upstream of its translation initiation codon (ATG). A PSI-BLAST search shows that this putative protein is highly homologous to terminal proteins that have been found in five linear plasmids and four linear chromosomes in *Streptomyces* (Table 3.6.2-1), so it was predicted to have a similar function as a terminal protein and designated as TpgCL1. These terminal proteins have very similar sizes and highly conserved amino acid sequences (Fig. 3.6.2-1). No conserved domain was detected in TpgCL1 on searching the NCBI-CDD database. However, InterPro Scan predicted a helix-turn-helix (HTH) motif, situated toward the N-terminus of TpgCL1. The characteristic DNA

Fig. 3.6.1-5. A comparison of colony growth indicating the transformation efficiency of four recombinants.

Colonies of pSG207, pSG207P, pST1 (the positive control) and pSG230 (the negative control) were patched on an MYM plate containing 60 μg/ml neomycin. The plate was incubated at 28°C for 60 hours. The colonies used as inocula all came from R5 plates containing 60 μg/ml neomycin except for the pSG230 colony, which came from an R5 plate without neomycin.

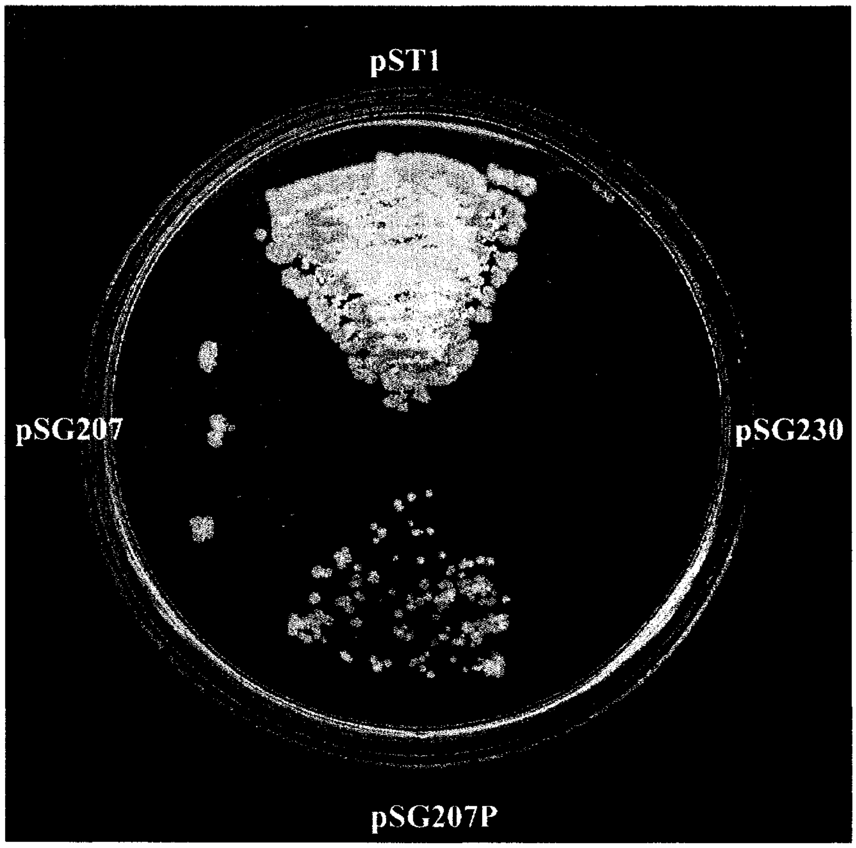

pST1
pSG207
pSG230
pSG207P

Table 3.6.2-1. Terminal proteins and telomere-associated proteins found in *Streptomyces* linear replicons.

Protein	Location	Species	Size (aa.)	Identity[a] (%)	Access Number
TpgCL1	pSCL2	*S. clavuligerus*	185	-	
TpgA1	chromosome	*S. avermitilis*	185	72 (134/184)[b]	gi:29611235
TpgA2	SAP1	*S. avermitilis*	185	75 (138/184)	gi:29611262
Putative terminal protein	pSV2	*S. violaceoruber*	185	73 (136/184)	gi:28797319
TpgC	chromosome, (identical to TpgL)	*S. coelicolor*	185	71 (131/184)	gi:15638581
TpgL	chromosome, (identical to TpgC)	*S. lividans*	185	71 (131/184)	gi:15638583
TpgSLP2	SLP2	*S. lividans*	184	62 (113/181)	gi:23428389
TpgR1	chromosome	*S. rochei*	185	51 (96/186)	gi:15638589
TpgR2	pSLA2-M	*S. rochei*	184	54 (98/181)	gi:15638585
TpgR3	pSLA2-L	*S. rochei*	184	54 (99/181)	gi:15638587
Putative terminal protein	pSV2	*S. violaceoruber*	176	43 (70/161)	gi:28797339
TapCL1	pSCL2	*S. clavuligerus*	220..256[c]	-	
TpgAR1	chromosome	*S. avermitilis*	752	57 (136/236)[d] 87 (223/255)[d]	gi:29611234
TpgAR2	SAP1	*S. avermitilis*	772	53 (130/241) 86 (221/255)	gi:29611263
Tap$_L$	chromosome, (identical to Tap$_C$)	*S. lividans*	739	54 (120/222) 85 (220/256)	gi:29424110
Tap$_C$	chromosome, (identical to Tap$_L$)	*S. coelicolor*	739	54 (120/222) 85 (220/256)	gi:29424112
Putative transcriptional regulator	pSV2	*S. violaceoruber*	656	61 (86/139) 89 (229/256)	gi:28797318
TapR1	chromosome, (identical to TapR3)	*S. rochei*	738	62 (146/235) 58 (144/248)	gi:29424108
TapR2	pSLA2-M	*S. rochei*	734	68 (152/223) 58 (143/246)	gi:29424106
TapR3	pSLA2-L, (identical to TapR1)	*S. rochei*	738	62 (146/235) 58 (144/248)	gi:29424108

a. Identity to the TpgCL1 or TapCL1 protein in pSCL2.

b. The fractions in parentheses represent the number of identical amino acid over their alignment region

c. Size of the N-terminal and C-terminal parts of the TapCL1, respectively

d. Identity to the N-terminal (upper) and C-terminal (bottom) of the TapCL1, respectively.

Fig. 3.6.2-1. Alignment of the amino acid sequences of Tpg proteins in *Streptomyces* linear plasmids.

The amino acid sequences of the Tpg proteins of linear plasmids of *S. clavuligerus* (TpgCL1), *S. avermitilis* (TpgA2), *S. violaceoruber* (pSV2.82), *S. lividans* (TpgSLP2) and *S. rochei* (TpgR2 and TpgR3), and of the chromosomes of *S. avermitilis* (TpgA1), *S. coelicolor* (TpgC), *S. lividans* (TpgL) and *S. rochei* (TpgR1) are aligned. The consensus of these sequences is also shown. The residues shaded in black are 100% conserved within the above sequences, and other residues shaded in gray are highly conserved as well.

```
TpgCL1(pSCL2)   MSSEFGGGLDTAVEKAFTRPAPKAAGTRMRYLVKHLKGTKAVAELLGVSQRTVERYVKA
TpgA2(SAP1)     M-SLFGDGLDAAVQKAFTRPAPKSAGAQMRYLVKQLKGTKAVAQMLRISQRTVERYVKD
pSV2.82         M-SLFGDGLEAAVQKAFTRPAPKSAGAQMRYLVKQLKGTKAVAQMLRISQRTVERYVKD
TpgA1           M-SMFGDGLEAAVHKAFTRPAPKSAGTQMRYLVKQYKGTKAVAQLLRISQRTVERYVKD
TpgSLP2(SLP2)   M-GIIGDGLDRAVQGAFTRPIPKSAGAQMRYLVKQLKGTRAVAGLLGVSQRTVERYVKD
TpgC, TpgL      M-SLFGNGLDAAVQKAFTRPAPKSAGAQMRYLVKQLKGTKAVAQMLRVSQRTVERYVKN
TpgR3(pSLA2-L)  M-DSIGDGLDRALESAFTRRPPQSAQAQMKYLVKQLKGTRAVARLLRISQRTVERYVSG
TpgR1           M-DSIGDGLDRALESAFTRRPPQSAQAQMKYLVKQLKGTRAVARLLRISQRTVERYVAG
TpgR2(pSLA2-M)  M-DSLGDSLDRALEGAFTRRIPQSAQAQMKYLVKQLKGTKATAQALGISQRTVERYVSG
Consensus       M-SLFGDGLDAAVQKAFTRPAPKSAGAQMRYLVKQLKGTKAVAQMLRISQRTVERYVKD

TpgCL1(pSCL2)   RS-KTRPDLAARLEREVKARWQPQIKAKARKKAATTGGIILDIHARMGYTAPIGTTDQDR
TpgA2(SAP1)     QIKKPRPDLAARLEREVKARWQPQIRAKAREKAATTGGIVVDTRARLGYTAPIGSTDQDR
pSV2.82         QIKKPRPDLAARLEREVKRRWQPQIRAKAKERAATTGGIVIDTRARLGYTAPIGSTDQDR
TpgA1           QIKKPRPDLAARLEREVKKRWQPQIRAKARQQAATTGGIVIDTRARLGYTAPIGSTDQDR
TpgSLP2(SLP2)   QIRRPRADLAQRLEDAVRQRWQPRVRDRARKQAAASTGLVIHTRARFGFTAAPGTTDDAR
TpgC, TpgL      EIKRPRPDLAARLEREVKARWQPQVRARARQKAATTDGIVIDTRARLGYTAPIGSTDQDR
TpgR3(pSLA2-L)  KLKRPRQDLRGRIEREVKKRWQPQVRAKARKKAASTDGLVVSTRARFGFTAAPGTTDDAR
TpgR1           QLKRPRRELRDRIEREVHKRWQPQVRARARRRAATTDGLVVSTRARFGFTAAPGTTDDAR
TpgR2(pSLA2-M)  KLKRPRQDLRGRIEREVKKRWQPQVRAKARKKAASTDGLVVSTRARFGFTAAPGTTDDAR
Consensus       QIKKPRPDLAARLEREVKARWQPQIRAKAREKAATTGGIVVDTRARLGYTAPIGSTDQDR

TpgCL1(pSCL2)   IRHITVALPPRHAARLLTAQDQGAGEDRLRELTAEALKETYFQDNGRRAGSLEEVKINDV
TpgA2(SAP1)     IRHLTVALPPVYAARLFDAQEAGASDARLQEIAAEALKEVYFQDGGRRAGSLEEVRFTDI
pSV2.82         IRHLTVALPPQHAARLFDAQEAGATEQRLQELAAEALKEVYFQDGGRRAGSLEEVRFTDI
TpgA1           IRHLTIALPPRYAAQLFEAQEQGASDQQLQEIAAEALKEVYFQDGGRRAGSLEEVRFTDI
TpgSLP2(SLP2)   IRHLTLALPPHHAARLLDAQDAGASEQQLRGLAAEALGEVYFRDGGRRAGGL-EVEFTDV
TpgC, TpgL      IRHLTVALPPQYAGRLFDAHQAGATDQQLQGIAAEALKEVYFQDGGRRAGSLEEVRFTDI
TpgR3(pSLA2-L)  IRDITQALPPEYADRLFTAREQGATEHQLQQIAADGLAQMYFRANDTRAHGL-GVEFTDI
TpgR1           IRDIIQALPPEFAARLFTAREQGATEHQLQQIAADGLAQMYFRANDTRAHGL-GVEFTDI
TpgR2(pSLA2-M)  IRDITQALPPEYADRLFTAREQGATEHQLQQIAADGLAQMYFRANDTRAHGL-GVEFTDI
Consensus       IRHLTVALPPVYAARLFDAQEAGASDARLQEIAAEALKEVYFQDGGRRAGSLEEVRFTDI

TpgCL1(pSCL2)   VHLDFEL
TpgA2(SAP1)     EHLEFEL
pSV2.82         EHLEFDL
TpgA1           EHLEFDL
TpgSLP2(SLP2)   EQAEFDL
TpgC, TpgL      EHLEFDL
TpgR3(pSLA2-L)  EQIEIQL
TpgR1           EQIEIQL
TpgR2(pSLA2-M)  EQIEIQL
Consensus       EHLEFEL
```

binding activity of the HTH motif matches the function of terminal proteins. This large family of HTH DNA binding proteins also includes bacterial plasmid copy control proteins, bacterial DNA methylases and various bacteriophage transcription control proteins. A similar HTH motif was also found in other *Streptomyces* terminal proteins (Bao & Cohen, 2001).

ORF 8.25.8 and 9.69.1 are two incomplete ORFs located at the right end of contig25 and left end of contig69, respectively. A PSI-BLAST search revealed that these two ORFs encode two different parts, the beginning 220 amino acids and the final 256 amino acids, of the same protein, which is highly homologous to the telomere-associated proteins encoded within various linear plasmids and chromosomes in *Streptomyces* (Table 3.6.2-1). The putative ribosome binding site (GGAGG) is located 8-bp upstreamof its translation initiation codon (GTG). The protein encoded by these two ORFs was thus predicted to be a telomere-associated protein as well and designated as TapCL1, encoded by the *tapCL1* gene. The *tapCL1* gene is located immediately (13 bp apart) upstream of the putative terminal protein gene, *tpgCL1*, of pSCL2, which also agrees with the gene arrangements of *tpg* and *tap* genes in *S. rochei*, *S. lividans* and *S. coelicolor* (Bao & Cohen, 2003).

Compared to the Tpg proteins, the Tap proteins have a lower level of sequence similarity among various species. Their sequence similarity is even lower in the region close their N-termini (Table 3.6.2-1).

As is found in other Tap proteins (Bao & Cohen, 2003), a predicted helix-turn-helix DNA binding motif (HTH_3 family domain) was located close to the N-terminus (28-83 amino acid) of TapCL1 in pSCL2, as detected by analysis using the Simple Modular Architecture Research Tool (SMART).

3.6.3 Other Proteins Involved in Replication

Besides the Rep and the terminal proteins that are directly involved in the replication process in linear plasmids, partitioning proteins for the replicon segregation process are also required, especially with low-copy-number plasmids. This section presents the identification of the partitioning proteins and helicases, which are also important for plasmid replication.

3.6.3.1 Partitioning proteins

The complete ORF 1.68.2 found near the left end of the linear plasmid encodes a 296 amino acid protein with a theoretical molecular weight of 32.6 kDa and a pI of 5.37. The putative RBS (GGAGG) was found to be located 6 bp upstream of the translation initiation codon, ATG. After two iterations of a PSI-BLAST search it was found that this putative protein has low (around 20%) but significant homology to partitioning proteins in many bacterial species, including the ParA proteins in *Corynebacterium*, *Streptomyces*, *Sinorhizobium*, *Agrobacterium*, *Brucella*, *Bradyrhizobium*, *Mesorhizobium*, *Enterococcus*, *Deinococcus*, *Bacillus*, soj proteins in *B. subtilis*, *S. coelicolor*, and SpoOJ regulator in *Treponema pallidum*. However a DNA binding domain and/or an ATPase-involved domain, which are usually found in other partitioning proteins, was not detected in this protein.

ORF 1.68.2a is located immediately downstream of ORF 1.68.2 (46 bp apart), and encodes a 10.3 kDa protein with 100 amino acids. Although no significant similarity was found between this putative protein and the known proteins in GenBank after several iterations of a PSI-BLAST search, its size and the genetic organization of these two ORFs implied that it could possibly have a ParB function. Previous studies have shown that ParB proteins are generally much more diverse than ParA proteins, and a meaningful alignment of the ParB sequences is possible only between closely related species (Gerdes, *et al.*, 2000). Based on their sizes and the conserved domain search result, the putative ParA and ParB proteins in pSCL2 are likely to belong to the Ib group (Gerdes, *et al.*, 2000), whereas partitioning proteins in other *Streptomyces* plasmids, including SCP1, SAP1, pSV2 and SLP2, all seem to be of the Ia type. These partitioning proteins will be compared and discussed further in the next chapter.

It is known that many type Ib loci have multiple direct repeats both upstream and downstream of the *par* genes. This feature was also found flanking the *par* locus of pSCL2. The upstream repeats may act as the centromere-like region, the *parS* sites, to be bound by the ParB protein. Three copies of 22-bp imperfect direct-repeats/palindromes (C(G/C)CCGTTTCCAA(C/T)GGAAACG) were found in the region 634-600 bp upstream of the *parA* gene. Two copies of 11-bp perfect direct repeats (CTTGGAGAG) were found 85-400 bp downstream of the *parB* gene. The remaining shorter repeats in this region are shown in Fig. 3.6.3-1.

3.6.3.2 Helicases

ORF 2.26.1 and ORF 2.26.6 both encode putative helicases. The sequences of the 5'-ends of these two ORFs have not been determined yet, so their physical characteristics cannot be analyzed at this stage. The translation product of ORF 2.26.1 (303 aa) is 75% identical to a 294 aa C-terminal fragment of the helicase encoded on the *S. avermitilis* linear plasmid SAP1 (SAP1_95, 885 aa), 71% identical to a 289 aa C-terminal fragment of the helicase encoded on the *S. violaceoruber* linear plasmid pSV2 (pSV2.02, 867 aa.) and is similar to many other *Streptomyces* putative helicases. Three and one half copies of conserved helicase associated domains (HA, 72 residues) were detected. This domain is predicted to contain three alpha helices and may function to bind nucleic acid.

The product of ORF 2.26.6 (774 amino acids) has low (around 30% identity) but significant similarity to helicases in many prokaryotic and eukaryotic species, from *Streptomyces* to human. The 100 amino acids in the centre of this protein are 76.8% aligned with a DEAD/H helicase superfamily C-terminal domain HELICc (82 residues), and 44.5% similar to the superfamily II helicase.

ORF 7.207.3c is within the gapped-contig207, consisting of the 3'-end of contig207-I and 5'-end of contig207-II with a gap (about 1 kb) between them. The N-terminus and C-terminus of its product have 30% and 40% identities to the putative SNF2/RAD54 family helicase (962 aa) in *S. avermitilis* and a putative helicase (890 aa) in *S. coelicolor.*

Fig. 3.6.3-1. Direct repeats located upstream and downstream of the *par* genes.

Regions highlighted by the same colour represent repeat sequences. The repeats downstream of the *parB* gene are underlined to distinguish them from the upstream sequences. The putative RBS (the italic, underlined letters), start and stop codons (bold letters) of *parA* and *parB* genes are shown.

1 TCGTGCGCGAGGGCATGACCCCCTCAGCAGCCCACGCCCAGGTTCTGGCCCGCCAGCGGC

61 AGCCCGAAGAAAACCGTGGCGTCTCCGCCGTTTCCAATGGAAACGCCAACGAGACCACCG

121 CAGAGGGCGTCCCCGGCCCGGCGCACCCCGCCGCCCCCGTTTCCAACGGAAACGGCTCCA

181 CCGGCAGCTCCGCGGCTCCGATCGCCGTTTCCAACGGAAACGCCGCCCC[redacted]

241 [redacted]CAGAGGAAGC[redacted]CCGAGG[redacted]AGGACGCCGACCGGCGTGCCGCAG

301 CCACTGCCCGG[redacted]GACTTCTGCTGGAAGCGGCCGACATCTCCAGCCCCG

361 ACACCCACGACTTGGTCATCAATGTGCTGACCGCCGCCGTCCTGGCCC[redacted]AGG

421 T[redacted]TCAGCAGCGCGCCTTCACCTGGCTGCGCGACACGGCCCGTCACGGTCTGG

481 ACGCCGCAGACGCTTCGGCGTACTTCAGCGCTGTGCAGGAGTCGGGTGACTCCACCGTCC

541 AGCGCCTCGCGGCCTTCGCCAGCGCCCTGGCGACGGGGGAGCTTCGGACCGCAGCGCGAC

601 GCCAGTCCTGGGGCACCCGCGAGATCCAGCACGTCCGCGTCCTCCAGCAGCACGCCCAGT

661 ACGACCCGCAGACCGAGTGGGAGAAGAAGGAACTCGGTATGGTCGCCGCA*GGAGG*AGACC

RBS

721 G**ATG**AGCATCAACTCGCGCTTCCGTCCCAAGGGACGTAACTGGCCCCTCATCTGG ......

parA →

parA |

1561 GGACTACCGCAAGGTCATGACCCACCTCGGCTTCCCGGACCTGGCAGCC**TGA**CCGTTTCC

1621 AACGGAGACGGGGGACAGGAAGCGTCTCGACTTCGGC**ATG**AGTCCAGGGCAAATCG ......

parB →

parB |

1921 AGCAGCCCGGCCCGGCCCCGGTCAGGGCCTCCAGCTG**TGA**CACGCCGCGGA[redacted]

1981 [redacted]ACGGAGGACGAAGTCACCGCGCCAGATGCCGGGATGAGACCGGCTCCCGGCCTCCGG

2041 GCCCCTTGGAGAGAGACGAGGGTGCGGGGGAGCGACCCACCGAGTACGGAGGCCCCCATG

2101 GTCGACCCGATCCCGACGCACATCTCGTCCGAGCGGCTGGAGCGCCTGGCGGAATTCGCC

2161 GACGCAGCCCGGCTGCGCGCGGCGCTGTACCCGTTGCCGAAGC[redacted]CGGCGC

2221 CGCTCCCGCCCCCGGAAGTACCCCGGTGGCACGCCGAACCCGTGGCGCACCGCGGTGGGC

2281 CACCGCCGGGGATGAGCGGCCACGCCGGACCTCACCGTGCACTCGGTTCGGGCCAGCGGT

2341 GGAGTCCACTTGGAGAGTGGAGACCATGGTGGTGCGCATCGAGTGGACTCCATTCCGGTG

2401 TGTT[redacted]GGAGTACGTTGCATCGTGCGCAACTCAGCGCACCGTGCGCGCGC

2461 CGGTCCGCACCTGAGCGCCTCCGCCTCGGTCATCAGGCGCTCGTGGCACG[redacted]

2521 [redacted]AGACCTCGGCTCTGCGATCCGCGCGGGGGACAGACCCCGGCACAGAAGGCCGCGCAGGC

3.6.4 Repeat Proteins

One complete open reading frame (ORF 3.73.5) in contig73 was found to encode a protein with highly conserved sequence repeats. This protein is 1009 amino acids in length. A 400 amino acid segment (593 - 995 aa) near the C-terminus of this protein consists of nine and half tandem repeat units, with 42 amino acids in each unit. A comparison of these repeat units shows that they can be perfectly aligned and are highly conserved (Fig. 3.6.4-1). The calculated molecular weight of this protein is about 109.5 kDa, with a predicted pI (isoelectric point) of 5.27. The putative ribosome binding site (RBS), GGAGA, is located 5 bp upstream of the start codon (GTG) of this ORF. This distance is within the normal range in *Streptomyces*. A BLAST search of this protein's sequence shows 29% identity (44% similarity) to a putative ATP/GTP binding protein in *S. coelicolor*, and 25% identity (43% similarity) to a putative ATP/GTP binding protein in *S. avermitilis*, respectively.

One of the most striking features of the protein encoded by ORF 3.73.5 is the length (42 amino acids) of its repeat unit. It is very reminiscent of a WD-40 protein, which typically contains 4-16 tandem repeat units, and characteristically about 40 amino acids ending with the conserved residues tryptophan (W) and aspartic acid (D) in each repeat (van der Voorn & Ploegh, 1992; Smith *et al.*, 1999). Members of the WD-family are found in all eukaryotes, serving as regulatory proteins with diverse functions (van der Voorn & Ploegh, 1992; Smith *et al.*, 1999). Recently WD proteins were also found in *Thermomonospora* (Janda *et al.*, 1996) and *Streptomyces* (Stoytcheva *et al.*, 2000). The best-characterized WD-repeat protein is Gβ subunit of heterotrimeric G proteins, which have a beta-propeller structure, where each blade of the propeller consists of four β strands (Fig. 3.6.4-2A). Fig. 3.6.4-2B shows structural elements within a single repeat in a WD protein, including four β strands connected by loops and turns (Smith *et al.*, 1999).

In order to compare the structures of the unknown repeat protein encoded by pSCL2 and the WD repeat proteins, the secondary structure of the hypothetical protein was predicted by the PepTool program. This structural analysis suggests that the last 34 amino acids in each repeat contain two alpha helixes (Fig. 3.6.4-1), and the entire repeat region consists of 72% alpha-helix but 0% beta-sheet. This result does not correspond to

Fig. 3.6.4-1. Amino acid sequence of the hypothetical protein encoded by ORF 3.73.5 and the consensus of the highly conserved repeat sequences in the protein.

The total length of this protein is 1009 amino acids. The sequence in regular letters, including amino acids 593 to 995, illustrates nine and a half repeat units. Each repeat unit contains 42 amino acids. The small italic letters represents sequence that does not contain a repeat. The bold letters show the consensus of the above repeats. The locations of two alpha helices in the last 34 amino acids of each repeat unit are shown at the bottom.

VGPVRWGQPGGVAPLQFEGAGFPAFAADTGDQAHFEYLRGELADVSTSSSGWVLDCPVWPAASPDSERPWAGASGAAVFCDGRLVGVVAEDNRAMGFRRLHAVPVH
EALSLPGFADLVTRHGHPGTTALPEEVTAGSGRTKSAEDNGMWPVEVGPVPALASAFQPRHMLRDRIDTAKARNGTVVLAQVLSGGGGVGKTQLAAACATDAFREG
VDLVVWALATEAQTVITRYAQAATRLQLPGVSGEDPPTDARLLLDWLATTRRRWLVVLDDLTDPAALGAWWPVSRTGTGWVLATTRMKGPMGGGRARVDIDVYEPA
ESAAYLRERLHGENMGHLLDDQAPTLADALGHLPLALGLAAAYMANEELTCAAYLRRFTDRRTRLNEALPDWADMEGYGRQITTALLLSLDAARATDTTGLAEPVL
RLTALLDPAGHPHTLWTTQTLLNHLTTTHTTPQAPQITADHTHSALRLLHRYALLTCDTRAEPRAVRIHALTARAVRENVSDPVLPELAGTAGDALLEIWPEADQP
HADLAAVLRANTDALAHHTGDHLWHPEGHTVLYHAGRSLLNAGLHGPAITYWQHMTERAEQL

```
LGDEHPGTLTARANLASSYWQAGRTEKAITILEQVVEDRQRL              (593-995 aa)
LGDEHPGTLTARANLAVSYWQAGRTDEAITIEEKVVEDRQRL
LGDEHPDTHTARANLASSYWQAGRTEKAITILEQVVEDRQRL
LGDEHPDTLTARANLAASYQQAGRTEKAITILEQVVEDRQRL
LGDEHPDTHTARANLASSYWQAGRTEKAITILEQVVEDRQRL
LGDEHPDTLTARANLAVSYQQAGRTDEAITILEQVVEDIERL
LGDEHPGTLTARANLASSYWQAGRTDEAITIEEKVVEDRQRL
LGDEHPGTLTARANLAASYQQAGRTEKAITILEQVVEDRQRL
LGDEHPGTLTARANLAVSYQQAGRTDEAITILEQVVEDRQRL
LGDEHPGTLTARANLAASYQQAGRT                                   LPGYADGPMNGTDG


LGDEHPGTLTARANLASSYWQAGRTEKAITILEQVVEDRQRL              Consensus
      D H       V  Q     DE    E K     IE
                A
```

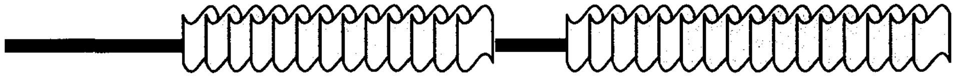


Fig. 3.6.4-2. The WD repeat.

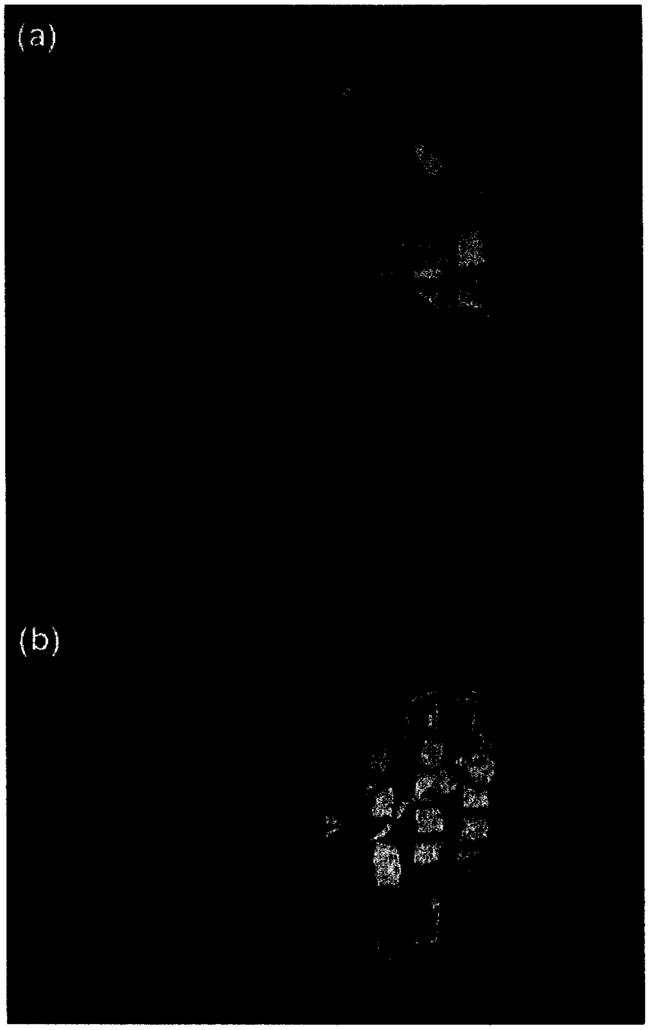
A. Structure of a WD-Repeat Protein.

The thin and gray lines show the backbone α-carbon trace of Gβ (based on the coordinates of Brookhaven Protein DataBank entry 1gp2). **(a)** Top view. **(b)** Side view. Each blade (one of which is shown in green) of the propeller consists of a single four-stranded antiparallel β-sheet. The N-terminal and C-terminal WD repeats are shown in red and yellow, respectively.

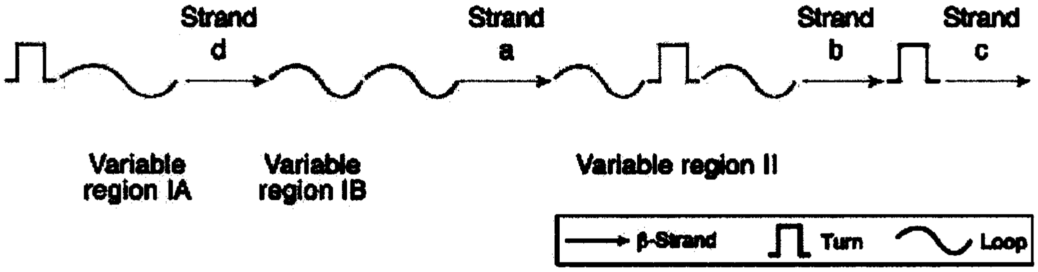
B. Schematic illustration of the structural elements within a single repeat. The four β-strands are illustrated as wide ribbons in each blade in the panel A. The loops and turns between the β-strands are illustrated as thin and coloured lines.

This figure was adapted from Smith *et al.*, 1999.

A.


(a)
(b)

B.


Strand d
Strand a
Strand b
Strand c
Variable region IA
Variable region IB
Variable region II
β-Strand
Turn
Loop

the characteristic structure of the proteins in the WD family so that eliminates the possibility of classifying this repeat protein into the WD family.

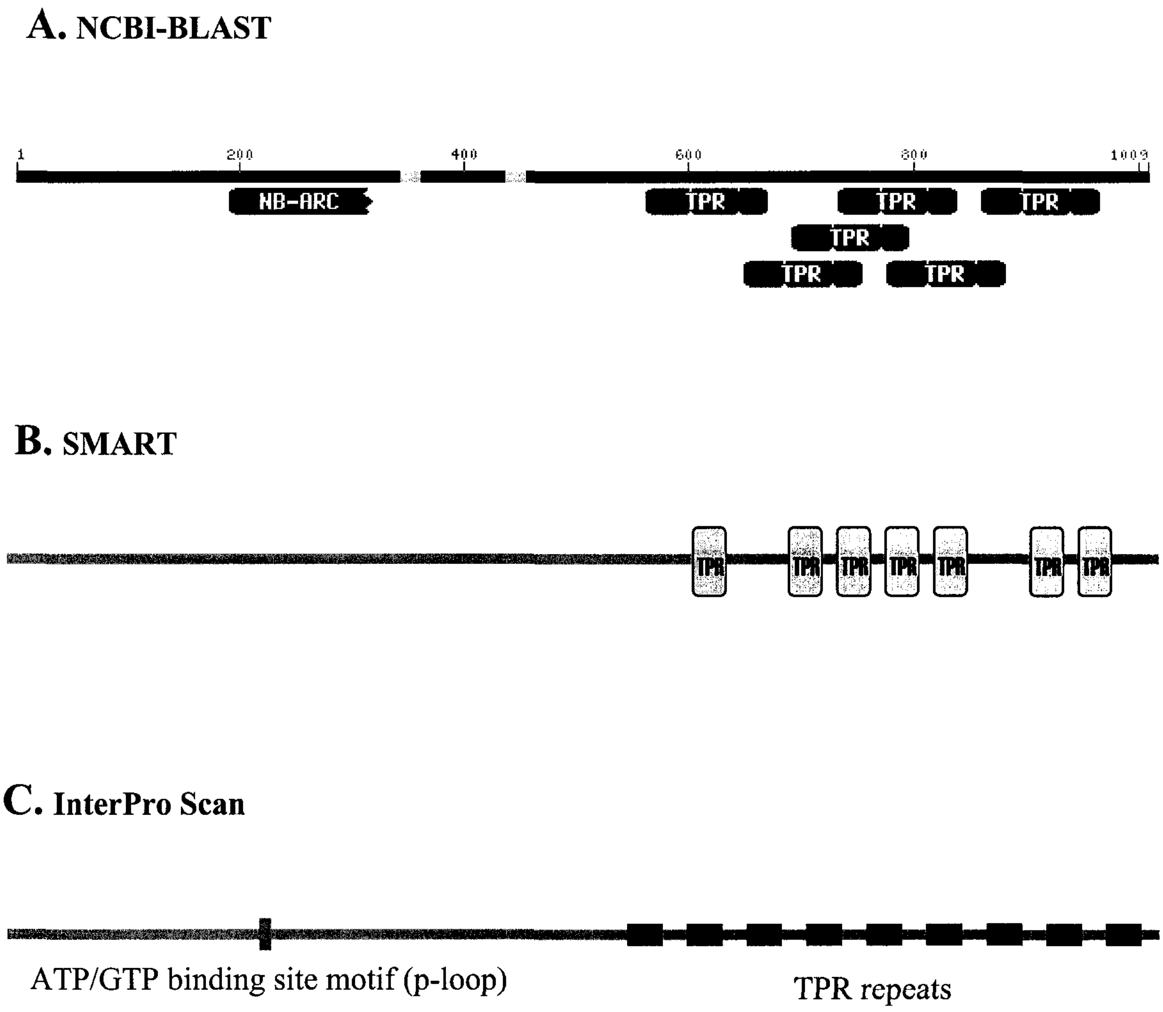
A search of this repeat protein for conserved domains using the BLAST-CD program found one regulation domain, an NB-ARC domain, at the N-terminus (Fig. 3.6.4-3A), which provides further evidence that this is probably a regulatory protein. The program also detected six TPR (Tetratrico Peptide Repeat) domains in the repeat region of this protein (Fig. 3.6.4-3A). A TPR domain is a degenerate 34-amino acid repeated motif and protein-protein interaction module found in many species with various functions, including cell cycle control, protein transport, and signal transduction. (Blatch & Lassle, 1999; Lamb *et al.*, 1995). A typical TPR protein normally contains 3-16 of these TPR units. Previous studies of secondary and tertiary structure showed that tandem arrays of TPR domains generate a right-handed helical structure (Fig. 3.6.4-4A), and a single TPR motif contains a pair of antiparallel alpha-helixes (Fig. 3.6.4-4B), which justmatch the two predicted alpha helixes in each repeat unit of the pSCL2 repeat protein. The 34 amino acids composing the two alpha helixes also match the size of the TPR motif. Further motif searches were done using the online programs InterPro Scan and SMART (Fig. 3.6.4-3B, C). The InterPro Scan prediction indicates nine TPR motifs at the C-terminus and an ATP/GTP binding site close to the N-terminal end, which conforms to the structure of the pSCL2 repeat protein most closely.

On the basis of structural analyses and homology searches, it appears that the protein encoded by ORF 3.73.5 is very likely to be a tetratrico peptide repeat protein, with an ATP/GTP binding site. It contains multiple copies of TPR motifs organized in a tandem array and each copy contains a pair of alpha-helices. However, compared to other proteins in the TPR family, the repeats in the pSCL2 repeat protein are much more highly conserved. In addition, the sequences connecting these TPR motifs in the protein encoded by pSCL2 are also conserved, which is different from other TPR proteins.

ORF 3.73.5 is not the only gene encoding a repeat protein in pSCL2. The product of ORF 4.H9.9c (309 amino acids; the N-terminal sequence has not been obtained) contains 7 copies of a highly conserved repeat. As seen with the repeats in the protein encoded by ORF 3.73.5, there are also 42 amino acids in each repeat unit of this protein.

Fig. 3.6.4-3. Tetratrico Peptide Repeat (TPR) motifs and other conserved domains in the pSCL2 repeat protein.

The repeat sequences in the protein encoded by ORF 3.73.5 were analyzed using three different programs, CD-search in NCBI-BLAST (A), SMART (B) and InterPro Scan (C).



A. NCBI-BLAST

B. SMART

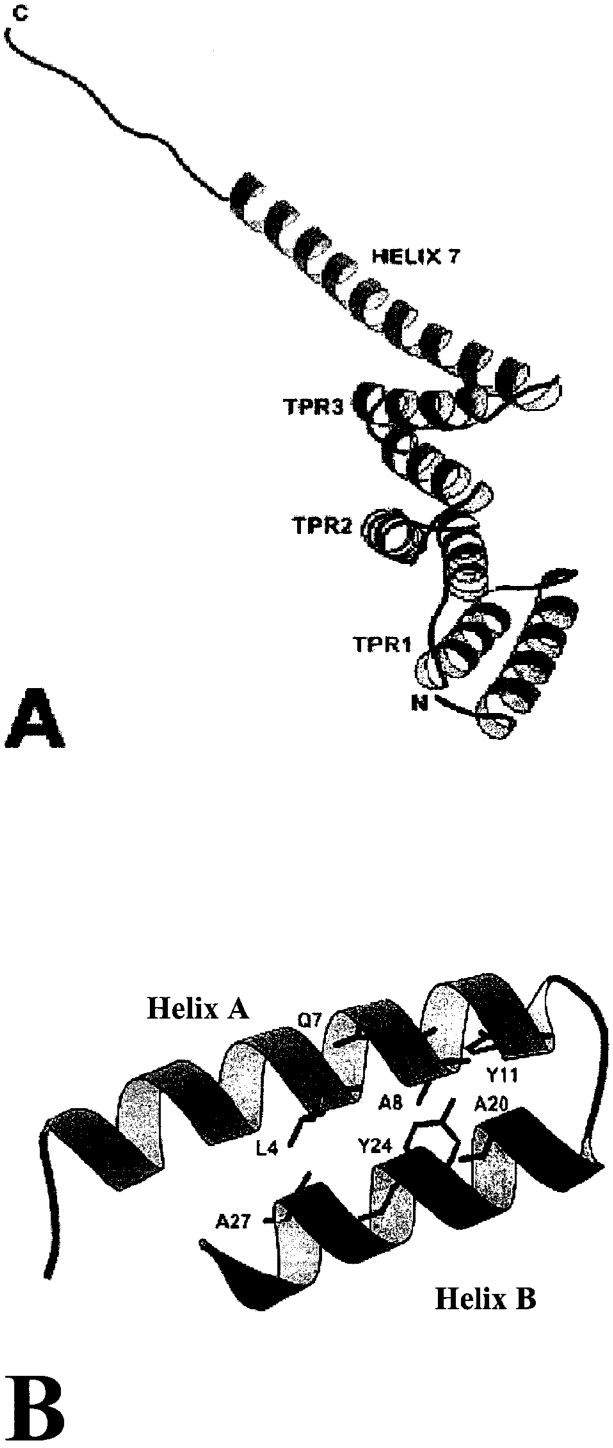
C. InterPro Scan

Fig. 3.6.4-4. Ribbon representation of the TPR domain of the human protein phosphatase 5.

A. View perpendicular to the helical axis of the TPR domain to illustrate the right-handed helical structure generated by tandem arrays of TPR motifs

B. Side view of a single TPR motif composed of a pair of antiparallel alpha-helixes, Helix A and Helix B. The highly conserved TPR consensus residues are labeled using the single-letter code and the numbering refers to the position of these residues in a motif.

This figure was adapted from Blatch & Lassle, 1999.


C
HELIX 7
TPR3
TPR2
TPR1
N
A
Helix A
Q7
Y11
A8
A20
L4
Y24
A27
Helix B
B

An alignment of these repeats and the consensus obtained is compared with the consensus from the protein encoded by ORF 3.73.5 (Fig. 3.6.4-5). The identity between these two repeat proteins in pSCL2 is around 60% (188 identical amino acids / 309 amino acids in the alignment region). A BLAST search of this incomplete protein sequence shows 41% identity (54% similarity) to a putative ATP/GTP binding protein in *S. coelicolor*, and 41% identity (56% similarity) to a putative serine/threonine protein kinase in *S. avermitilis*. The high similarity between the repeat regions in these two pSCL2-encoded repeat proteins suggests that these repeats may play an important role, probable as recognition sites, in functioning of the regulatory proteins.

3.6.5 Transposases / Integrases in Insertion Sequences

Two ORFs encoding putative transposases were found in pSCL2, which led to the discovery of two insertion sequences (IS) in this linear plasmid.

The first insertion sequence, 1425 bp in length, is 99.8% identical to IS*116* that was first found in *S. clavligerus* by Leskiw *et al.* (1990) and suggested to be located in the chromosome, at a time when the two large linear plasmids, pSCL2 and pSCL3, had not yet been discovered. The hybridization results in the same study also indicated that only one copy of IS*116* was present per genome of *S. clavuligerus*. This, together with the identity between the IS found in this work and IS*116*, suggests that these are likely to be the same insertion sequence and located on the linear plasmid, pSCL2, instead of the chromosome in *S. clavuligerus*. Unlike most other insertion sequences, the termini of IS*116* do not contain inverted repeat (IR) sequences, which usually characterize the ends of IS elements (Iida *et al.*, 1983).

IS*116* encodes a putative transposase/integrase (400 amino acids), with a calculated molecular weight of 44.8 kDa. Its high isoelectric point (10.14) is a typical feature of transposases. Two conserved transposase domains are located on this transposase (Fig. 3.6.5-1): the sequence from 75-172 aa is 100% aligned with the transposase_9 domain (52% identity), which is found in the transposases of the IS*111A*/IS*1328*/IS*1533* family; and the sequence from 246-355 aa is 100% aligned with

Fig. 3.6.4-5. Amino acid sequence of the C-terminus of the protein encoded by ORF 4.H9.9c.

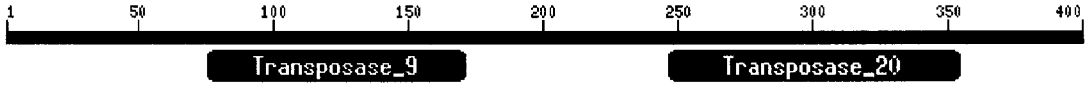
The highly conserved repeats in the protein are aligned, and the close relationship to the repeat units in the protein encoded by ORF 3.73.5 is shown. The “consensus 1” and “consensus 2” in italic letters represent the consensus of the repeats from the proteins encoded by ORF 4.H9.9c and ORF 3.73.5, respectively. The bold letters show the alignment of the two consensus sequences.

```
   ...      TNLAASYQQAGRTEEAIELLEQVLTASERV
LGPNHPDTLTTRTNLAASYQQAGRTEEAIELLEQVLTASERV
LGPNHPDTLTARGNLAGSYQQAGRTEEAIELEEQVLTGRDRI
LGPDHPKTLTTRGNLAFSHWQAGRTEEAIELEEQVLTARDRI
LGPNHPDTLTARGNLAASYQQAGRTEEAIELLEQVLTARERV
LGPDHPDTLTARGNLAFSHWQAGRTEEAIELEEQVLTARDRI
LGPNHPDTLTTRGNLAASYQQAGRTEEAIELLEQVLTASERV
LGPDHPRTIVTR                               RALERWRTEPGGEEA

LGPNHPDTLTTRGNLAASYQQAGRTEEAIELLEQVLTARERV       consensus 1 (ORF 4.H9.9c)
   D  K  VA T   F HW           E     GSD I
      R         G

LGDEHPGTLTARANLASSYWQAGRTEKAITILEQVVEDRQRL       consensus 2 (ORF 3.73.5)
      D H       V  Q     DE    E K    IE
                A

LG--HP-TLTAR-NLA-SYWQAGRTEEAI--LEQV---RER-       Alignment of consensus 1 and 2
        HVT       HQ      DK    E K    IQ
                                       SD
```

Fig. 3.6.5-1. Conserved domain search for the transposases in IS*116* and IS*Scl1*.

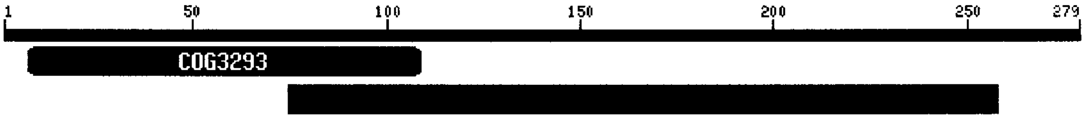
The coloured boxes under the line represent the detected domains. The sequences below each diagram show the alignment between the query sequence and the searched domain sequence from the database. The red and blue letters indicate the identical and similar residues. The conserved carboxylate residues in the DDE superfamily are highlighted. The total lengths of the conserved domains and the percentage of their total lengths aligned are also shown.

IS*116*



```
CD-Length = 99 residues, 100.0% aligned

    IS116:       75    LVAHGQNVVYVPGRTVNRMSGAYKGEGKTDAKDARVIADQARMR-RDFAPLD  125
    Transpo_9:   1     LGAAGLKVVYVNPRAIARFAKALGGRAKTDAKDACVIARYARTGLHRLRPLL  52

    IS116:       126   RPPELVTTLRLLTNHRADLIADRVRLINRLRDLLTGICPALERAFDY  172
    Transpo_9:   53    PPDDIVVELRELTRRREDLVATRTRLANRLRRLLREYFPAAERAFDS  99

CD-Length = 110 residues, 100.0% aligned

    IS116:        246  LLALDERIKDNDREIRETFRTDDRAEIIESMPGMGPVLGAEFVAIVGDLSGYKDA  300
    Transpo_20:   1    LRELDEQIKDIDAEIEELLRLHADAQILRSIPGVGPITAATLLAEIGDPSRFKSA  55

    IS116:        301  GRLASHAGLAPVPRDSGRRTGNYHRPQRYNRRLRWLFYMSAQAAMMRPGPSRDYY  355
    Transpo_20:   56   RQLAAYAGLAPRPRSSGRKTGRGGISKRGNRRLRRALYMGALVALRHDPGSRAFY  110
```

IS*Scl1*



```
CD-Length = 124 residues, 80.6% aligned

    IS468':  7     YPSDLSDARWALIEPTLTAWRNARLERRPTGQPAKVELRD--VFNAILYLNRTGIPWKYL  64
    COG3293: 2     KLHALVDAEWRPVEPLLP-------PAKYGGPPGVTLLRDREVLNGIADLLYTGCAWRAL  54

    IS468':  65    PHDFPGHGTVYFYYAA--WRDEGIFAQLNYDLTALARVKEGRKPDPT  109
    COG3293: 55    PADFPPATTVIPYRRFRRWFKRGLWKRRNLVERTFGRLKQFRRTATR  101

CD-Length = 231 residues, 98.3% aligned

    IS468':  74    VYFYYAAWRDEGIFAQLNYDLTALARVKEGRKPDPTASVIDTQSVKTSTNVPLTSQGTDA  133
    DDE:     1     LLRRFRRLESVSLLRELLRRLLASLLTAKLGGSGRSVLIIDSTTVRTPGSPEGKGKK-KG  59

    IS468':  134   AKKICGRKRGILTDTI-GLILAVTVAGAGLSENAMGIRLLDQAKASYPTIAKSWVDTGF-  191
    DDE:     60    KRGYGGVKLHIAVDTGTGLPLALVVTPGNVHDSAVLEALLDLLDE-DPKGVLVLADAGYD  118

    IS468':  192   KNAVIEHGATLGID----------------------------------------VEVVNR  211
    DDE:     119   GAELLEKLEEKGVDYLIRVKKNAKLKDKKKTLKKLKRRRVLARGETKGEREKEYRYVTNL  178

    IS468':  212   NPEIRGFHVVKRRWVVERSIGWIMLHRRLARDYETLTASSEAMIRIASI  260
    DDE:     179   PEAEEVAELYRLRWQIERVFKWLKRFFGLRRLRERSFNRAEAELLLALL  227
```

transposase_20 domain (48% identity), which belongs to the transposases of IS*116*/IS*110*/IS*902* family. The transposase encoded in IS*116* has a high degree of similarity to many transposases encoded by other prokaryotic IS elements as well, especially in high GC content Gram-positive species. This similarity is obvious when its sequence is aligned with other transposase sequences. The closest homologs are listed in Table 3.6.5-1.

The second insertion sequence found in pSCL2 was named IS*Scl1*. It is 1097 bp long, of which the last 988 bp are 88% identical to the last 988 bp of IS*468A/B* found in *Streptomyces coelicolor*. IS*468A* and *-B* are identical in nucleotide sequence and are tandemly aligned with a common direct repeat (DR; CTCAG) between them (Yamasaki *et al.*, 2000). IS*Scl1* and IS*468A/B* have exactly the same length, target sites (DR), and the same 9 bp perfect inverted repeat (IR) at both ends. The GC content of IS*468A/B* and IS*Scl1* are 62.2% and 62.8%, respectively, which is significantly lower than the GC content of IS*116* (69%) as well as the average GC of *Streptomyces* DNA (70-74%; Kieser *et al.*, 2000).) This suggests that these ISs could have been transferred relatively recently to *Streptomyces* so that they have not been fully affected by the pressure for high GC content in *Streptomyces* species (Yamasaki *et al.*, 2000).

The ORF in IS*Scl1* encodes a 31.3-kDa protein with 279 amino acids and a calculated pI of 9.62. It has 94.5% identity and 97% similarity to the transposase (279 aa) encoded by IS*468A/B*. The closest homologs of the protein encoded by IS*Scl1* are listed in Table 3.6.5-2. Two transposase domains were detected in this protein (Fig. 3.6.5-1): the sequence between amino acids 7 and 109 has 40% similarity to the COG3293 domain; and the amino acids 74 to 260 of this protein are 98.3% aligned with the transposase DDE domain. Although the sequence similarity between the IS*Scl1* transposase and DDE domain is low, the carboxylate residues, D and E (Asp and Glu), which are characteristic of the members in the DDE superfamily, are conserved in this transposase at three separate positions (114D, 118D, and 228E). Based on the description in the Conserved Domain Database of NCBI, these three carboxylate residues are believed to be responsible for coordinating metal ions needed for the catalytic activity of the transposase that involves DNA cleavage at a specific site followed by a strand

Table 3.6.5-1. Most similar homologs of IS*116* transposase.

Species	Function	Length (aa)	Identities %	Positives %
S. fradiae	Integrase	400	65	77
Rhodococcus rhodochrous	Putative transposase	400	60	74
S. avermitilis, SAP_1	Putative IS*116*/IS*110*/IS*902*-family transposase	397	52	69
Mycobacterium avium subsp. avium	P44 protein	399	52	67
M. avium subsp. paratuberculosis	Transposase	399	51	69
M. avium	IS*901* protein	401	51	67
M. avium	IS*902* protein	400	51	67
S. avermitilis	IS*110* family transposase	401, 396	51, 50	65, 64
M. avium subsp. paratuberculosis	Putative transposase	406	50	68
M.paratuberculosis	IS*900* hypothetical protein	399	50	68
M. avium	IS*1110* transposase	464	50	64
E. coli	Putative IS*110* transposase	398	34	52
S. coelicolor A3(2)	IS*110* transposase/integrase	405	34	49
Corynebacterium efficiens YS-314	Insertion element conserved hypothetical protein	427	33	50

Table 3.6.5-2. Most similar homologs of IS*Scl1* transposase.

Species	Function	Length (aa)	Identities %	Positives %
S. coelicolor A3(2)	IS*468* putative transposase	279	94	97
Methanosarcina acetivorans str. C2A	Transposase	278	40	60
Gluconacetobacter xylinus	Transposase	278	40	57
S. lividans	Putative transposase	320	40	57
S. avermitilis	Putative transposase	433	39	57
Synechocystis sp. PCC 6803	Transposase	261	38	57
Acetobacter xylinus	IS*1031A* probable transposase	278	38	56
Sinorhizobium meliloti	Putative transposase	277	38	55
Streptomyces coelicolor A3(2)	Putative Tn5714 transposase	281	37	54

transfer reaction. The DDE superfamily also contains transposases encoded in IS*4*, IS*421*, IS*5377*, IS*427*, IS*402*, IS*1355*, and IS*5*.

3.6.6 Transfer Proteins

Four ORFs in pSCL2 were found to encode analogs of the transfer protein Tra3 that was originally discovered in the 43-kb linear plasmid, pBL1, in *Streptomyces bambergiensis*. The Tra3 protein is one of the main components required for transfer of the linear plasmid pBL1 of *S. bambergiensis*, and for pock formation (Zotchev & Schrempf, 1994). The protein encoded by ORF 3.73.2 contains 287 aa, and is 45% identical to the Tra3 protein in pBL1. Its C-terminal and N-terminal fragments also have similarities to two putative Tra3 homologs, SAP1_34 (46% identity) and SAP1_32 (73% identity), in the 94-kb linear plasmid SAP1 of *Streptomyces avermitilis*, respectively. The rest of three Tra3 homologs in pSCL2 can be divided into two groups based on their sequence similarities. The product of ORF 6.A8.5 is homologous to the N-terminus (amino acids 1 to 135) of the Tra3 protein of pBL1 and the putative Tra3 homolog, SAP1_34, in SAP1. In the second group, the proteins encoded by ORF 6.A8.10 and the first 88 amino acids of the ORF 7.207.13 protein are homologous to amino acids 139 to 232 of the Tra3 protein of pBL1, putative Tra3 homolog SAP1_32 in SAP1, as well as a hypothetical protein in the linear plasmid pSV2 of *S. violaceoruber*. No significant similarity was found between the two groups. The properties of these homologous proteins are listed in Table 3.6.6-1A, and the identities among them are summarized in Table 3.6.6-1B. The amino acid sequence alignments of these Tra3 homologs are shown in Fig. 3.6.6-1. The existence of four Tra3 homologs in pSCL2 suggests that this linear plasmid is transmissible.

In addition to the Tra3-like proteins, an analog of the TraA-like transfer protein has been found in pSCL2. ORF 2.26.13c encodes a protein (606 aa) that is 80% identical to SAP1_75 (608 aa) encoded by the linear plasmid SAP1 in *S. avermitilis*. The N-terminus (amino acids 4 to 271) and C-terminus (amino acids 273 to 606) of the ORF 2.26.13c protein also have 84% and 75% identity to a putative transfer protein (pSV2.28c, 269 aa) and a hypothetical protein (pSV2.27c, 333 aa) encoded by the linear

Table 3.6.6-1. Properties of the Tra3 protein and its homologs in pSCL2 and other *Streptomyces* linear plasmids (see the text for details).

A. Location and size of the Tra3 protein and its homologs

Tra 3 proteins and homologs	Location	Size (aa)	Tra 3 homologs	Location	Size (aa)
Tra3	*S. bambergiensis* pBL1	339	3.73.2	*S. clavuligerus* pSCL2	287
pSV2.76	*S. violaceoruber* pSV2	236	6.A8.5	〃	104
Tra3 homolog SAP1_32	*S. avermitilis* SAP1	110	6.A8.10	〃	159
Tra3 homolog SAP1_34	*S. avermitilis* SAP1	161	7.207.13	〃	181

B. Identities between the Tra3 proteins and the Tra3 homologs

			Group 1	Group 2	
		3.73.2	6.A8.5	6.A8.10	7.207.13
	3.73.2	-	-	NS	NS
Group 1	6.A8.5	37	-	NS	NS
Group 2	6.A8.10	NS[a]	NS	-	-
	7.207.13	NS	NS	52	-
Tra3 in pBL1	1-135 aa	45	54	NS	NS
	139-232 aa		NS	55	65
Tra3 homologs in SAP1	SAP1_32	73	NS	70	61
	SAP1_34	46	55	NS	NS
	pSV2.76	48	NS	55	48

a. "NS" indicates no significant similarity.

Fig. 3.6.6-1. Alignments of amino acid sequences of Tra3 protein and Tra3 homologs.

Tra3 proteins is the transfer protein in pBL1 from *S. bambergiensis*; 3.73.2, 6.A8.5, 6.A8.10 and 7.207.13 are Tra3 homologs in pSCL2 from *S. clavuligerus*; SAP1_34 and SAP1_32 are the Tra3 homologs in SAP1 from *S. avermitilis*; pSV2.76 is the Tra3 homolog in pSV2 from *S. violaceoruber*. Black shaded "*" symbols signify identical residues. Grey shaded "." or ":" symbols represent conserved or highly conserved residues, respectively. The alignment was carried out using the CLUSTAL W (1.82) multiple sequence alignment program. The similarities between these proteins are shown in Table 3.6.6-1.

```
Tra3    1 M-SGG----------ETSGVDLARVALRAAMKAARKNGSGQKAKQKQRPVRTV-RRDGRE 48
3.73.2  1 M-TETITTVGEQPALELSGVDLARVALHQAREAARARGESGTRKAKRRPLTMAGRRDGRE 59
SAP1_34 1 MMTETP---------QLSGVDLARVALRAAKVAAQKNGGGRTAQPKPR-TTRVVRRDGRE 50
6.A8.5  1 ----------------VSG---SMSTRFRGNPPASWTAQTVTEPSFSRGTTTVVRRDGRE 41
                         .:**:::::.:.  .  .*       .     *  : . ******

Tra3      PMGLGAAIGALVTE-------------------------------LAVSYDPDSGRLTVC 77
3.73.2    PTGFAAVLQSLVADRAWDVPTAGGSILDQWPAIAVAVSPNLPAHVTAVAFHPDTGQLDLR 119
SAP1_34   PMGLGAAISALVTERAWELPAAGASLRARWEAIAPDFG-----HVVAVGCDADSGRLTVC 105
6.A8.5    PLGLGAAIGMMMTER------------------------------GLAAPSTGG----- 64
          * *:.*.:  ::::::.... .. ..   . ...         :: .:.  .. * :

Tra3      PESAAWATKARLEQTRVIAAANEAAGRAVVRALRILPPGAVPAPSPADVAPEMPADAPTG 137
3.73.2    PDSSAYATQLRLIGSRIVTTANNTTGTQAVRTVRVLAVGAASTPHCEPAAAPTPATA--A 177
SAP1_34   PESAAWATKLRLEQARVVEAANESAGRTVVRGLRILAPGSVPVSEPADVTPES--GCP-- 161
6.A8.5    PEAVAGVRGARERVGKWLRPGFRVRGS---------------------------------  91
          *:: * .:  *.   . .  ..   *   :: .:.:  :... . ......  .. .

Tra3          PVR----TRETACEGYRRALAAHQQVAVPSRVDPGIAEAVERQTAAMRELSRRRV 188
3.73.2        PVK----TRETASAGFHQALAAHQEAAPPSRVDPHIAEAVERQTKAMRELSRRAF 228
pSV2.76   128 PVK----TREMACDGYRRALAAHQTVRPDRHVDPAIQEAIESQTRALRELSQRAF 178
SAP1_32    1  -MR----TRETASDDYRRALAAHQEVAPPRRVDPAIAEAVERQTAAMRELSRRAF  50
6.A8.10    39 VVRRYRPARRTPPNGYCRAIEAHRHAARTSRVDPAVAEAVERQTKATRTLSRRAF  93
7.207.13   1  ----------MGSASFHQALAAPQAVVPPSRVDPSIAEAVERQSAAMRALSLLAF  45
              .::    :*...  .: :*:.*:: . ....*** :.**:*.*: * *.**.*::

Tra3     SEADVVAP-DDALASIEATRAQRRRQAAATEAAALRRARSEKAGC  232
3.73.2   PEPD-AAP-DDTAVPIE----------------------------  243
pSV2.76  P--DLAS--GDQPSPIEAARVQPRRDAAASRATALRRARAERAQ-  218
SAP1_32  PEPDVVS--DDGPAPIEQARAQRRRQAAATEAAALRRARQERA--   83
6.A8.10  PEPDAVA--DEEPAPIDQARVQRRRQAAATEAAALRRARAEGAG-  136
7.207.13 PEAQEPADYAPASDRAGPHRAPPPSCGDRSRRPSPRPGRARPAER   90
         .:.: .:  ..   .:. .* ::::.::::.:.::*::* : *
```

plasmid pSV2 in *S. violaceoruber*, respectively. It seems likely that there could be a sequence error(s) at the end of pSV2.28c, so pSV2.28c and pSV2.27c are very likely to be one ORF encoding a transfer protein. Based on the above sequence analysis, ORF 2.26.13c is assumed to encode a TraA-like transfer protein with a theoretical molecular weight of 66 kDa and a pI of 10.44.

Unlike the Tra3 homologs that only have similarities to proteins encoded by plasmids from *Streptomyces*, the TraA-like protein in pSCL2 has homology (42 to 49% similarity) to many TraA proteins in other actinomycetes, including *Rhodococcus* and *Corynebacterium*. This TraA-like protein also has distant homology (33 to 38% similarity) to TraI, TrwC, helicase, relaxase or *ori*T nicking/unwinding proteins in plasmid F, R100, R46, and in some Gram-negative species, such as *E. coli.*, *Shigella*, *Salmonella*, and *Novosphingobium*, which suggests that this transfer protein may have helicase or relaxase activities (Llosa *et al.*, 1994). The top alignments in some of the homology categories are shown in Fig. 3.6.6-2.

3.6.7 Regulatory Proteins

The regulatory proteins of *S. coelicolor* were subdivided into 13 families in a report of the *S. coelicolor* genome project (www.sanger.ac.uk/Projects/S_coelicolor/ scheme.shtml). Genes representing three of these families, GntR, LysR and MarR, have been found in pSCL2.

A PSI-BLAST search for proteins having homologies to the putative protein encoded by ORF 4.H9.4 (244 aa, Mw 26.5 kDa, pI 6.73) of pSCL2 shows that it has low but significant similarity to many GntR-family transcriptional regulators in *Streptomyces*. The homologs of the ORF 4.H9.4 protein include a putative GntR-family regulatory protein (252 aa) in *S. coelicolor* (29% identity and 43% similarity), and many Kor proteins. Kor proteins are known to regulate the process of replication, transfer, integration and excision in *Streptomyces* plasmids. For example, in the self-transmissible circular plasmid pIJ101, the *korA* and *korB* genes regulate the transcription of the *tra* and

Fig. 3.6.6-2. The sequence alignments between pSCL2 TraA-like protein and its homologs.

The similarity searches and the sequence alignments were carried out using PSI-BLAST with two iterations. "Query" represents the pSCL2 TraA-like protein, while "Sbjct" represents its homologs. "-" represents gaps.

Putative TraA-like protein [SAP1, *Streptomyces avermitilis*]

```
 Length = 608, Score = 762 bits (1969), Expect = 0.0
 Identities = 486/613 (79%), Positives = 513/613 (83%), Gaps = 12/613 (1%)

Query: 1   MISMSPVRPGSGWRYLFRGVMAGDGHRAPGTSLRAAQDEAGVPPGVWKGGGLAAVGLAAG 60
           MISMS VRPG+GWRYLFRGVM GDGHR  G SLRAAQDEAGVPPGVWKG GLAAVGL AG
Sbjct: 1   MISMSEVRPGNGWRYLFRGVMVGDGHRPAGKSLRAAQDEAGVPPGVWKGRGLAAVGLKAG 60

Query: 61  DVVTERQAELLLGEGRHPDADRIERELLDAGHDPAAARRAAVLGRPIEHNRSPETEKAKE 120
           DVVTERQAELLLGEGRHPDADRIERE L  G  PA ARRA VLGRPIEHN+SP+T+KAKE
Sbjct: 61  DVVTERQAELLLGEGRHPDADRIERERLAEGKSPAQARRATVLGRPIEHNQSPKTDKAKE 120

Query: 121 RIPWLAFDLVFRPPSTAHIAWALMDDETRRVLEECQDTARDKTLAWLEESVAQIRWGSGG 180
           R PWL FDLVFRPP TAHIAWAL D ETR VLE CQD ARDKTLAWL E VA+IRW  GG
Sbjct: 121 RTPWLGFDLVFRPPPTAHIAWALGDYETRLVLELCQDIARDKTLAWLGEEVAEIRWKGGG 180

Query: 181 KHRKPVRDGLIVTVFRHYESRAGQ--PLLHDHAVVSIRARRPDD-GTWGNLSADSLLANI 237
           KHR  VRDGLIV VFRHYESRA +  PLLHDHAVVSIRARRPDD GTWGNLSADSL+A+I
Sbjct: 181 KHRARVRDGLIVAVFRHYESRAAESKPLLHDHAVVSIRARRPDDKGTWGNLSADSLMAHI 240

Query: 238 VAADTLYTLHFMEEVSARLGWAWEPREVTSGRRPVMEIAGIDQRLIGWQSTRRQQIEDAL 297
           VAADTLYTL+FMEEVSARLGWAWEPREVT GRRPVME+AGIDQRLIGWQSTRRQQIEDAL
Sbjct: 241 VAADTLYTLYFMEEVSARLGWAWEPREVTPGRRPVMELAGIDQRLIGWQSTRRQQIEDAL 300

Query: 298 PVLTARYEERQGHPPGERATYQLACQAADQTRPPKRTEPLSLNGLRARWRTSAIAAFGAC 357
            VLTA YE++QGHPPGERA Y L CQAADQTR PKRTE LSL  LR RWR SAI A+G
Sbjct: 301 SVLTANYEKKQGHPPGERAGYALGCQAADQTRSPKRTELLSLTELRERWRDSAIRAYGVD 360

Query: 358 TVYRLAQRARAAAAAVWARVKPAVDIALAAVDTVAVVYVMRGAFARRHLLAEARRHLAHA 417
              RLA+RARAAAAAVWARV+P VDIALAAVD VAVVYVMRGAF R HLLAEARRHL++
Sbjct: 361 VFDRLAERARAAAAAVWARVRPVVDIALAAVDVVAVVYVMRGAFKRHHLLAEARRHLSYV 420

Query: 418 LGGRPHPPGLDEQIVQTVVDDYTRPVGRGRAMTADLRALYPHDIGEQAVLRPLTRNRSAP 477
           L GRPH PGLDEQIVQ VVDDYTRPVGRGR MTADLRALYP D  +QAVLRPLTR RSAP
Sbjct: 421 LRGRPHRPGLDEQIVQAVVDDYTRPVGRGRMMTADLRALYPRDTEDQAVLRPLTRKRSAP 480

Query: 478 PYERARLAGGALATRVRAVRRAERLNSRAGPGAIVVPAASPGSRPRPFRTGRKTGRLLEP 537
           PYERARLA GALA RV A RRAERL SR  P A+ VPAAS  S PRPFRTG K GRLLEP
Sbjct: 481 PYERARLAAGALAARVNAARRAERLGSRPRPYAVAVPAASR-SHPRPFRTGPKAGRLLEP 539

Query: 538 ETGVDVDSVEQTRRTLE----AAAARLRDGIRERAAVHGPRPQTAPAPVAGTPPHTEQPG 593
           ETG  VD+VEQTR+TLE      AA ++D  R R A HG R Q  PAP A  PPHT+QPG
Sbjct: 540 ETG--VDAVEQTRQTLEAAAAKVAATIQDSRRAREAAHGLRSQ--PAPAAVPPPHTQQPG 595

Query: 594 TQHTPGRTTGGIA 606
            QHTPGR++GG+A
Sbjct: 596 VQHTPGRSSGGVA 608
```

TrwC protein [*Escherichia coli* plasmid R388]

```
 Length = 966, Score = 270 bits (692), Expect = 4e-71
 Identities = 84/326 (25%), Positives = 125/326 (38%), Gaps = 56/326 (17%)

Query: 27  RAPGTSLRAAQDEAGV--PPGVWKGGGLAAVGLAAGDVVTERQAELLLGEGRHPDADRIE 84
           RA       A D          W+G G   +GL +G+V ++R  ELL G
Sbjct: 14  RAASYYEDGADDYYAKDGDASEWQGKGAEELGL-SGEVDSKRFRELLAGN---------- 62

Query: 85  RELLDAGHDPAAARRAAVLGRPIEHNRSPETEKAKERIPWLAFDLVFRPPSTAHIAWALM 144
                             +G      RS   + +KERI     DL F  P +  +   +
Sbjct: 63  ------------------IGEGHRIMRSATRQDSKERI---GLDLTFSAPKSVSLQALVA 101
```

```
Query: 145 DDETRRVLEECQDTARDKTLAWLEESVAQIRWGSGGKHRKPVRDGLIVTVFRHYESRAGQ 204
            D     + +  D A  +TL   E + AQ R    GK R      L++  FRH  SR
Sbjct: 102 GDAE---IIKAHDRAVARTLEQAE-ARAQARQKIQGKTRIETTGNLVIGKFRHETSRERD 157

Query: 205 PLLHDHAVVSIRARRPDDGTWGNLSADSLLANIVAADTLYTLHFMEEVSARLGWAWEPRE 264
           P LH HAV+    +R  DG W  L  D ++        +Y      E+  +LG+     +
Sbjct: 158 PQLHTHAVILNMTKR-SDGQWRALKNDEIVKRTRYLGAVYNAELAHELQ-KLGY-----Q 210

Query: 265 VTSGRRPVMEIAGIDQRLIGWQSTRRQQIEDALPVLTARYEERQGHPPGERATYQLACQA 324
           +  G+    ++A ID++ I   S R +QI        A +   +G  P   +  Q   QA
Sbjct: 211 LRYGKDGNFDLAHIDRQQIEGFSKRTEQI--------AEWYAARGLDPNSVSLEQK--QA 260

Query: 325 ADQTRPPKRTEPLSLNGLRARWRTSA 350
           A      K+T  +    LRA W+ +A
Sbjct: 261 AKVLSRAKKTS-VDREALRAEWQATA 285
```

Mobilization protein TraI [*Escherichia coli*]

```
 Length = 1078, Score = 262 bits (669), Expect = 2e-68
 Identities = 119/533 (22%), Positives = 182/533 (34%), Gaps = 95/533 (17%)

Query: 30  GTSLRAAQDEAGVPPG--VWKGGGLAAVGLAAGDVVTERQAELLLGEGRHPDADRIEREL 87
           G    A  D          W+G G  A+GL +GDV + R  ELL+GE
Sbjct: 17  GYYSDAKDDYYSKDSSFTSWQGTGAEALGL-SGDVESARFKELLVGE------------- 62

Query: 88  LDAGHDPAAARRAAVLGRPIEHNRSPETEKAKERIPWLAFDLVFRPPSTAHIAWALMDDE 147
                                H +    +  KER   L +DL F  P    +   +  D+
Sbjct: 63  ----------------IDTFTHMQRHVGDAKKER---LGYDLTFSAPKGVSMQALIHGDK 103

Query: 148 TRRVLEECQDTARDKTLAWLEESVAQIRWGSGGKHRKPVRDGLIVTVFRHYESRAGQPLL 207
           T   + E  + A    +   E+ +AQ R    GK      + L+V  FRH  SRA  P L
Sbjct: 104 T---IIEAHEKAVAAAVREAEK-LAQARTTRQGKSVTQNTNNLVVATFRHETSRALDPDL 159

Query: 208 HDHAVVSIRARRPDDGTWGNLSADSLLANIVAADTLYTLHFMEEVSARLGWAWEPREVTS 267
           H HA V    +R +DG W  L  D L+ N +    +Y      E++ + G+        +
Sbjct: 160 HTHAFVMNMTQR-EDGQWRALKNDELMRNKMHLGDVYKQELALELT-KAGYELR----YN 213

Query: 268 GRRPVMEIAGIDQRLIGWQSTRRQQIEDALPVLTARYEERQGHPPGERATYQLACQAADQ 327
            +    ++A      I   S R +QIE  L  +             E A  Q   + +
Sbjct: 214 SKNNTFDMAHFSDEQIRAFSRRSEQIEKGLAAMGLTR---------ETADAQTKSRVSMA 264

Query: 328 TRPPKRTEPLSLNGLRARW--RTSAIAAFGACTVYR------LAQRARAAAAAVWARVKP 379
           TR  K+TE  S   +   W  R   +        ++       A  AR  A    +
Sbjct: 265 TRE-KKTEH-SREEIHQEWASRAKTLGIDFDNREWQGHGKPLEADIARNMAPDFTSPEVK 322

Query: 380 AVDIALAAVDTVAVVYVMRGAFARRHLLAEARRHLAHALGGRPHPPGLDEQIVQTVVDDY 439
           A      AV +++       +F R+ L+  A + +            L    +  V   Y
Sbjct: 323 ADRAIQFAVKSLSE---RDASFERQKLIQIANKQV------------LGHATIADVEKAY 367

Query: 440 TRPVGRGRAMTADLRALYPHDIGEQAVLRPLTRNR-------SAPPYERARLAG------ 486
            + V +G  +  + R      +G   +   LTR         S    ++AR A
Sbjct: 368 LKAVQKGAIIEGEARYQSTLKVGASVMAETLTRKEWIDSLTNSGMRADKARFAVDDGIKN 427

Query: 487 GALATRVRAVRRAERLNSRAGPGAIVVPAASPGSRPRPFRTGRKTGRLLEPET 539
           G L      V   E    R     + + +   G  PR   T    G+LL  +T
Sbjct: 428 GRLKKTSHRVTTVE--GIRLERSILTIESRGRGQMPRQL-TAEIAGQLLSGKT 477
```

```
DNA helicase [Novosphingobium aromaticivorans]
 Length = 1013, Score = 339 bits (870), Expect = 9e-92
 Identities = 102/419 (24%), Positives = 153/419 (36%), Gaps = 64/419 (15%)

Query: 1    MISMSPVRPGSGWRYLFRGVMAGDGHRAPGTSLRAAQDEAGVPPGVWKGGGLAAVGLAAG 60
            M S++ VR  SG         A D          +A + A    GVW G G  A+GL
Sbjct: 1    MHSIASVRSSSG---------AADYFANDNYY--SADEHAE--AGVWGGEGARALGLEGQ 47

Query: 61   DVVTERQAELLLGEGRHPDADRIERELLDAGHDPAAARRAAVLGRPIEHNRSPETEKAKE 120
                ER A   +  GR PD + + +                     +E  R
Sbjct: 48   V---ERDAFEGVLNGRLPDGEMVGQ---------------------VEGRR--------- 74

Query: 121  RIPWLAFDLVFRPPSTAHIAWALMDDETRRVLEECQDTARDKTLAWLEESVAQIRWGSGG 180
                L  DL F  P +A I   +  D  RR+++      +      +E+  A+ R
Sbjct: 75   ----LGLDLTFSMPKSASILALVSGD--RRIIDAHLAAVKSTMSQLVEKQFAESRNYERS 128

Query: 181  KHRKPV-RDGLIVTVFRHYESRAGQPLLHDHAVVSIRARRPDDGTWGNLSADSLLANIVA 239
            +  +P     L+  +F H  SRA  P  H HAVV+   R P  GTW  L    +  N
Sbjct: 129  RSGEPQKTGNLVYALFAHDTSRALDPQGHIHAVVANLTRDP-KGTWKALWNGEIWKNNTT 187

Query: 240  ADTLYTLHFMEEVSARLGWAWEPREVTSGRRPVMEIAGIDQRLIGWQSTRRQQIEDALPV 299
                Y   F  ++  +LG+  E     +G+    EI G+   +I   STR  +IE  +
Sbjct: 188  IGQFYHAAFRAQLQ-KLGYETE----AAGKHGSFEIKGVPAEVIKAFSTRTTEIEAKIAE 242

Query: 300  LTARYEERQGHPPGERATYQLAC--QAADQTRPPKRTEPLSLNGLRARWRTSAIAAFGAC 357
            + A   E +          +LA   + A      +R   L  +G        A A   A
Sbjct: 243  VGATRLETKKQITLYTRDPKLAVEDRGALVEGWQQRAAELGFDGKALVAAAKARAEVEAR 302

Query: 358  TVYRLAQRARAAAAAVWARVKPAVDI-ALAAVDTVAVVYVMRGAFARRHLLAEARRHLA 415
              +R  + A AA   V  R+  A+   +  AV   A +++       +H  A A RHL+
Sbjct: 303  PSFR--ETATAAIGEVSTRINAALRTPSPLAVSGAAALFLSAATIKAQHATASAIRHLS 359
```

kilB genes, respectively, which are required for the efficient transfer of pIJ101 (Kendall & Cohen, 1987; Stein *et al.*, 1989). The putative ORF 4.H9.4 protein has similarities to the KorSA proteins of the chromosomes in *S. avermitilis* (29% identity and 42% similarity) and *S. coelicolor* (24% identity and 41% similarity), of the circular plasmid pSAM2 in *S. ambofaciens* (25% identity and 41% similarity), and to the predicted KorA protein (encoded by ORF-L) of pSCL1 in *S. clavuligerus* (21% identity and 35% similarity). In addition, it also shows remote homologies to members of the GntR family of transcriptional regulators in many other bacterial species, including *Bacillus, Enterococcus, Yersinia, Agrobacterium, Brucella, Streptococcus, Pseudomonas,* and *Ralstonia.* A conserved domain search showed that the ORF 4.H9.4 protein appears to be homologous to a transcriptional regulator domain PhnF, and the N-terminus (amino acids 14 to 73) of it has 43% identity and 53% similarity to a helix_turn_helix gluconate operon transcriptional repressor domain (HTH_GNTR) in the gntR family. This conserved domain search result further confirmed that this protein belongs to the GntR family. The putative ribosome-binding site, GGAGG, is located 5 bp upstream of its start codon, GTG.

The product of ORF 7.207.15 (311 aa, Mw 33.5 kDa, pI 8.70) was found to have 75% identity to a putative LysR-family transcriptional regulator in *S. avermitilis*. The conserved transcriptional regulator domain LyR was 100% aligned with this protein. The N-terminus of this protein also has homologies to LysR-type transcriptional regulators in *E. coli* (43% identity), *Shigella* (43% identity), *Salmonella* (42% identity), *Pseudomonas* (59% identity), *Ralstonia* (46% identity), and *Yersinia* (35% identity). This indicates that the regulatory proteins in this family are widely used in various bacterial species.

ORF 5.424.4 encodes a putative protein that is 160 amino acids in length. A BLAST-CD search showed that amino acids 67 to 149 of this putative protein are 100% aligned with the "Acetyltransf" domain (82 residues) in the Acetyltransferase (GNAT) family, which contains proteins with N-acetyltransferase functions. However the overall protein has 29% identity to a putative transcriptional regulator of the MarR family in *S. avermitilis*, and 48% identity to a hypothetical protein in *Magnetospirillum magnetotacticum.*

In addition to the genes encoding regulatory proteins described above, pSCL2 also encodes regulatory proteins of other groups. Among these are ATP/GTP binding proteins, a serine/threonine protein kinase, an RNA polymerase sigma factor, a two-component regulator, and others.

CHAPTER 4 DISCUSSION

The majority of the nucleotide sequence of the linear plasmid, pSCL2, from *Streptomyces clavuligerus* has been determined. 82, 728 base pairs were assembled into ten contigs and gapped-contigs, which contain 98 predicted protein-coding genes. The overall GC content of the obtained nucleotide sequence is 69.97%, which is in the range for *Streptomyces* DNA.

4.1 SEQUENCING

DNA fragments of pSCL2, generated by random fragmentation (nebulization), were cloned into the *E. coli* cloning vector pCR®4Blunt-TOPO® to construct sequencing libraries. The major cloning problem encountered in this project was that some regions of the pSCL2 sequence could not be cloned, neither in the large insert library nor in the one with small inserts, though the total length of the inserts in the two libraries provided ten-fold coverage of the length of pSCL2. This problem may have been caused by the high-level expression of genes whose products are toxic to the *E. coli* host (Kurland & Dong, 1996). This may be especially problematic when genes from Gram-positive bacteria are cloned into a Gram-negative host. Gram-positive bacteria usually use ribosomal binding sequences (RBS) that are close to the consensus sequence (GGAGGA) for efficient translation initiation, which is not necessarily the case in Gram-negative bacteria. The over-expression problem can be avoided, or at least reduced, by using a low-copy-number vector. Over-expression of toxic genes mainly affects the clones with large inserts. The cloning problems in a relatively small-insert library may be simply caused by the unclonable regions in the target DNA sequences. Previous studies have shown that cells carrying plasmids containing repeat sequences, long and continuous GC-rich sections, or with potentially complex secondary structures are less likely to survive.

On the contrary, it was also observed that some regions were cloned repeatedly. A number of clones from two independent libraries enclose exactly same inserts. These inserts are not necessarily within or adjacent to a palindromic region that causes DNA secondary structures and makes DNA strands easy to break. The probable cause for this problem may be contamination of a reagent with a vigorous clone.

Primer-walking was the major approach that was used in the later stages of the project to fill in gaps left by the random sequencing approach. Because primer-walking is a time-consuming method, it is not suitable for sequencing fragments larger than about 5 kb. The templates used were either clones from the sequencing libraries or PCR products amplified to cover unclonable regions. The design of suitable primers is critically important for successful sequencing reactions in primer-walking, as well as for obtaining pure, high yield PCR products when amplifying unclonable regions. All primers were designed to be between 18 and 22 nucleotides in length, with a melting temperature (T_m) between 58 and 68°C. A satisfactory primer should have a GC content as close to 50% as possible, especially at its 3' end. Efforts should be made to avoid the potential to form dimers, hairpins, long stretches of identical bases, or to have more than one G or C at the 3' end. Suitable primers should bind exclusively to a single site. It is not always possible to design an ideal primer for use with *Streptomyces* DNA since, as mentioned in previous chapters, *Streptomyces* DNA has an extremely high GC content and *Streptomyces* linear plasmids contain abundant direct and inverted repeats. The balance of the melting temperatures of the forward and reverse primers for PCR amplification is also important. Because *S. clavuligerus* genomic DNA was used as template in the PCR reactions, large T_m differences between the two primers might have caused false priming or inadequate annealing, which might have led false products or low yields.

The terminal regions of pSCL2, including the TIR, have not been included in the currently assembled sequence. Because the terminal 180-bp sequence has been published (Huang *et al.*, 1998), oligonucleotide probes were synthesized to screen the sequencing libraries in order to detect clones containing the end fragments of pSCL2. However, no positive hybridization signals were obtained. The use of short probes (about 20 nucleotides) could be one of the reasons for weak hybridization. However, it is virtually impossible to design a longer probe due to the abundant palindromes found in this region. It is also possible that the terminal sequence had not been cloned and included in the libraries. There could be several reasons why there might have been difficulties cloning the terminal sequences. As discussed above, numerous direct and inverted repeats in the terminal sequence make clones containing such inserts unstable in host cells, and may lead to death of the host cells. Alternatively, as a result of the linearity of the plasmid molecule it is possible that the terminal sequences may have a tendency to be left in larger fragments during the shearing process. If so, these larger fragments would be less likely to be cloned than those smaller ones. A third possibility is that the terminal-capped proteins were not eliminated completely so that one or more remaining amino acids esterified to the 5'-phosphates might have blocked ligation to the vector used for creating the library. To avoid such problems, it would be wise to clone the terminal sequence separately, using different cloning and sequencing strategies. On the other hand, the absent terminal sequence may not significantly affect the analysis of gene organization and function in pSCL2 since previous studies of other giant linear plasmids showed that few critical genes were found close to their ends.

4.2 ANNOTATION

DNA sequencing projects provide detailed sequence information, from which genes can be predicted and annotated, and the specified gene organization can be examined. In addition to pSCL2, five other giant linear plasmids from *Streptomyces* spp have been completely sequenced so far. The gene distributions of these linear plasmids are summarized in Table 4-1. The total length and the overall GC content of these linear plasmids are also shown. Other than the genes required for plasmid replication and segregation, each plasmid has its characteristic gene repertoire. Some of them have genes related to their known phenotypes, particularly those encoding biosynthetic enzymes for antibiotics.

Table 4-1. Summary of gene distribution of some *Streptomyces* linear plasmids.

Function *	SCP1	pSV2	SAP1	SLP2	pSLA2-L	pSCL2
Replication						
Replication proteins	2				6	2
Helicase	3	1	1	1	2	3
Partitioning proteins	4	2	2	2	2	2
Terminus(-associated) proteins		2	2	2	1	2
Biosynthetic proteins	21				66	
Degradation	1	1	2		1	
Macromolecule Synthesis	5		1	2	4	
Membrane associated	40	3			1	3
Metabolism of small molecules	12		1	2	5	1
Regulation	13	5	7	2	14	7
Spore associated	6			1		
Transfer proteins	3	3	5	6	3	5
Transposase	10	4	2	3	2	2
Other functions	26	6	10	5	7	6
Conserved hypothetical proteins	13	29	45	6	14	15
Hypothetical proteins	198	54	18	11	15	49
Total No. of putative proteins	357	110	96	43	143	97
Length of complete sequence (bp)	356,023	96,742	94,287	50,410	210,614	82,732
G+C content (%)	69.1	70.0	69.2	68.4	72.8	69.97

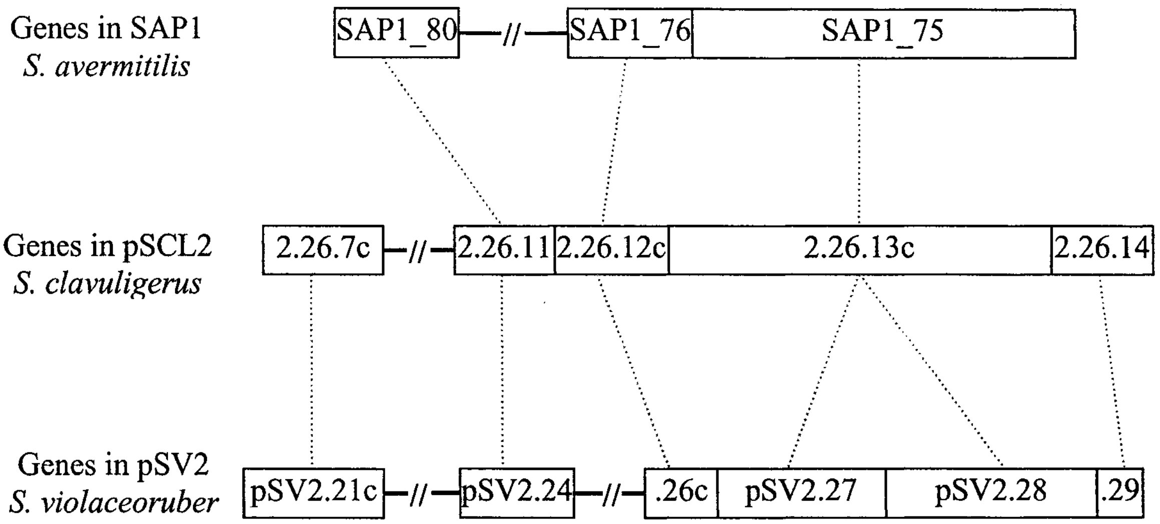
* Based on the protein function classification of the *S. coelicolor* genome suggested by Sanger Center (www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml).

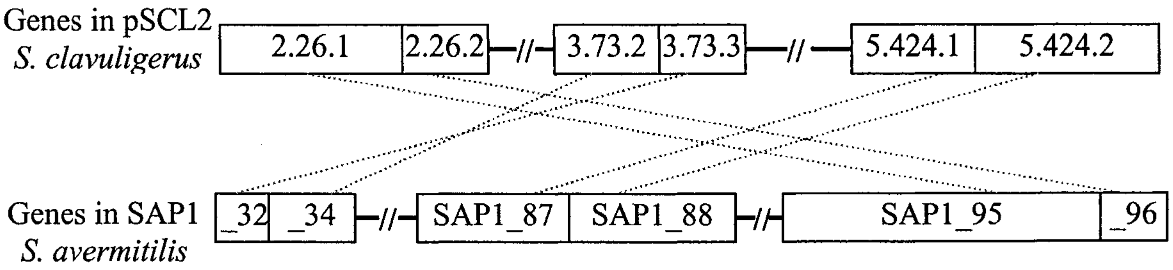
SCP1 of *S. coelicolor* carries a gene cluster for the biosynthesis of and resistance to the antibiotic, methylenomycin, as well as a large number of membrane associated proteins, which may facilitate secretion of secondary metabolites (Bentley *et al.*, 2002). The plasmid pSLA2-L of *S. rochei* contains even more condensed biosynthetic and regulatory proteins than SCP1 because it carries five biosynthetic gene clusters responsible for the synthesis of two antibiotics, lankacidin and lankamycin (Omura *et al.*, 2001). In addition, the finding of 10 genes encoding transposases on SCP1 indicates that it has a fair number of sequence elements capable of active transposition. No evidence has been found to indicate that the linear plasmids pSV2, SAP1, SLP1, and pSCL2 have any obvious relationship to secondary metabolite production. Almost half of the putative protein-coding genes on these plasmids encode hypothetical proteins with unknown functions. This is normal for plasmids since they do not have pressure to keep indispensable genes. Most of the hypothetical proteins in pSCL2 are clustered in the regions in contig4.H9, contig5.424 and contig8.25. Similar clustering can be observed in SAP1 and pSV2 too, but not as obviously as that in pSCL2. The similarities of pSCL2 to SAP1 and pSV2 are also reflected by their similar gene organizations in certain regions. For example, the gene products of 2.26.7c to 2.26.14 of pSCL2 resemble the proteins encoded by pSV2.21c to pSV2.29 in pSV2; among them, 2.26.11 to 2.26.13c also resemble SAP1_80 to SAP1_75 in SAP1, respectively(Fig. 4-1). More cases are found in several other regions, including 2.26.1-2 of pSCL2 to SAP1_95-96, 3.73.2-3 of pSCL2 to SAP1_34-32, and 5.424.1-2 to SAP1_87-88 (Fig. 4-1).

Most *Streptomyces* plasmids, including linear plasmids, contain an autonomous transfer mechanism (Kieser *et al.*, 2000). Transfer genes were found in all of the linear plasmids listed in Table 4-1. Among them, SLP2 and pSCL2 each encodes six transfer proteins. This is double the number of transfer proteins encoded in SCP1, which is six times and two times larger than SLP2 and pSCL2, respectively. SLP2 has been shown to be a self-transmissible plasmid and a sex factor in *S. lividans*, as it promotes both chromosomal recombination and the transfer of non-transmissible derivatives of other plasmids (Hopwood *et al.*, 1983). Recent studies have shown that it contains many genes for transfer and/or conjugation-associated proteins, which include a cluster of proteins for transfer-spreading, a transfer protein (TtrA) with helicase function, and a Kor protein

Fig. 4-1. Similar gene organizations in *Streptomyces* linear plasmids.

The boxes represent the predicted genes in each linear plasmid in *Streptomyces*. The grey broken lines connecting two genes demonstrate similarities between the two putative genes from two linear plasmids (see Table 3.5.1 for details).


Genes in SAP1
S. avermitilis
SAP1_80
SAP1_76
SAP1_75
Genes in pSCL2
S. clavuligerus
2.26.7c
2.26.11
2.26.12c
2.26.13c
2.26.14
Genes in pSV2
S. violaceoruber
pSV2.21c
pSV2.24
.26c
pSV2.27
pSV2.28
.29


Genes in pSCL2
S. clavuligerus
2.26.1
2.26.2
3.73.2
3.73.3
5.424.1
5.424.2
Genes in SAP1
S. avermitilis
_32
_34
SAP1_87
SAP1_88
SAP1_95
_96

(KorSLP2) for regulating transfer and/or replication of SLP2 directly (Huang *et al.*, 2003). The six transfer genes in pSCL2 are not clustered. These include one TraA-like protein and five homologs of the Tra3 protein that was originally found in pBL1 of *S. bambergiensis*. Tra3 and the other four putative genes in pBL1 are the only conjugal transfer related genes that have been identified and characterized in *Streptomyces*. The discovery of five Tra3 homologs in pSCL2 implies that pSCL2 may have similar conjugal transfer function as pBL1 does to facilitate the transfer of circular plasmids. Although there are no similarities between the transfer proteins of SLP2 and pSCL2, they may also share similar functions, including the ability to mediate conjugation and promote horizontal transfer of plasmids and donor chromosomes. Further experiments on these transfer genes can help us understand the function of this linear plasmid. A KorSA-like protein is also encoded in on pSCL2. It resembles the ORF-L in pSCL1 (Wu & Roy, 1993), Kor proteins in the circular conjugative plasmid pIJ101, and pSAM2 of *S. ambofaciens* (Sezonov *et al.*, 2000), and may regulate replication, transfer, integration and excision.

Like many plasmids, especially low-copy-number plasmids, putative genes encoding partitioning proteins, ParA and ParB, were found in pSCL2. However, their amino acid sequences do not show significant similarities to partitioning proteins encoded in other *Streptomyces* linear plasmids that have been sequenced so far. Table 4-2 compares the partitioning proteins in some of the linear plasmids. Based on their sizes and closest homologs, SCP1, SAP1, pSV2, SLP2 and pSLA2-L all seem to have type Ia partitioning loci (Gerdes *et al.*, 2000), which are highly homologous to each other. This is in contrast to the phylogram made by Gerdes *et al.* (2000), which showed that all type Ia partitioning loci were from plasmids of Gram-negative bacteria, and *par* loci from plasmids of Gram-positive origin were exclusively of the Ib type. The type Ia partitioning proteins of *Streptomyces* linear plasmids in Table 4-2 are similar to each other, but have relatively lower homologies to chromosome partitioning proteins of streptomycetes and other bacteria, which are smaller in size (about 300 to 350 aa). Based on Pfam search results, Gerdes *et al.* (2000) suggested that all the ParA proteins detected so far contain ParA family ATPase domains, and ParB proteins contain ParB-like nuclease domains. Examination of the chromosomal and type Ia plasmid partitioning proteins in Table 4-2

Table 4-2. Comparison of ParA and ParB proteins in the linear plasmids SCP1, SAP1, SLP2, pSV2, pSLA2-L and pSCL2.

	ParA			ParB		
	Size	**Motif**	**Closest Homologs***	**Size**	**Motif**	**Closest Homologs***
Type Ia						
SCP1	420 (ParA1)	soj	pSLA2-L.ParA (84) SCP1.ParA2 (55)	375 (ParB1)	SpoOJ ParB	pSLA2-L.ParB(75) SCP1.ParA2 (36)
	387 (ParA2)	soj	SAP1.ParA (58) SCP1.ParA1 (55)	381 (ParB2)	SpoOJ ParB	SAP1.ParB (54) SCP1.ParA1 (36)
SAP1	427	soj	SCP1.ParA2 (58) pSLA2-L.ParA (53)	343	SpoOJ ParB	SCP1.ParB2 (54) pSLA2-L.ParB(41)
SLP2	386	soj	SCP1.ParA2 (52) SCP1.ParA1 (46)	299	SpoOJ ParB	pSLA2-L.ParB(43) SCP1.ParB1 (40)
pSV2	385	soj	SCP1.ParA2 (57) SAP1.ParA (49)	278	None	SCP1.ParB2 (29)
pSLA2-L	419	soj	SCP1.ParA1 (84) SCP1.ParA1 (55)	409	SpoOJ ParB	SCP1.ParB1 (73) SCP1.ParB2 (38)
Type Ib						
pSCL2	296	None	*Corynebacterium glutamicum* ParA	100	None	No significant similarity

* Numbers in the brackets indicate the percentage of identity to ParA or ParB of pSCL2

showed that they all contain conserved soj motifs in their ParA, and most of them contain Spo0J and ParB motifs in their ParB. The previous homology search also showed that the similarity among ParA proteins is much stronger than the similarity among ParB proteins in different organisms.

Based on their sizes and the results of a BLAST search, the partitioning genes in pSCL2 probably belongs to type Ib, which matches to the phylogram results of Gerdes *et al.* (2000), but is inconsistent with the classification of *par* loci of other *Streptomyces* linear plasmids. The Par proteins encoded in pSCL2 also have similarities to chromosome partitioning proteins in various species, however, no conserved domain was detected on either ParA or ParB. Why does pSCL2 contain such unique partitioning proteins? Do these two putative genes encode real partitioning proteins of pSCL2? Further experiments and analysis need to be carried out to answer these questions.

The type Ib ParA protein is relatively smaller than the type Ia ParA proteins, and the ParB in type Ib is smaller still. In general, the ParB proteins are much more diverse than the ParA proteins, and a meaningful alignment of the ParB sequences is possible only within closely related species (Gerdes *et al.*, 2000). Even lower similarities were observed in type Ib ParB proteins. This proposition was further supported by the observations in this study. A BLAST search showed some similarity between the ParA protein encoded in pSCL2 and that in some *Corynebacterium* spp., but no significant similarity was found between the pSCL2 ParB and any known proteins. The type Ib ParA proteins encoded in the plasmids of *Corynebacterium diphtheria* (196 aa in pNG2), *C. efficiens* (195 aa in pCE2; 192 aa in pCE3), *C. glutamicum* (192 aa in pAG1; 199 aa in pTET3; 216 aa in pGA2), *C. jeikeium* (194 aa in pA501) and some other high GC Gram positive species including *Micrococcus* and *Bifidobacterium* are all highly homologous. However, the ParB proteins in these organisms do not share similarities.

4.3 Replication

Genes involved in the replication of these linear plasmids are compared in Table 4-1. All of these plasmids encode partitioning proteins, terminal proteins and helicases.

SCP1, pSLA2-L and pSCL2, but not SAP1, SLP2 and pSV2, encode replication proteins, which are essential for autonomous replication. This indicates that SAP1, SLP2 and pSV2 probably have to rely on chromosomal DNA polymerases or replicative proteins on other plasmids in their host for DNA synthesis. On the other hand, SCP1 has two ARS (autonomously replicating sequence), and pSLA2-L has three ARS (Huang *et al.*, 2003). ARSs of linear *Streptomyces* plasmids usually consist of an origin of replication and one or more genes encoding replication proteins that are generally located near the replication origin of the plasmid. So far, one ARS was found on pSCL2, in which the two replication genes, *repC1* and *repC2*, are highly homologous to the two replicative genes, *repL1* and *repL2*, located in ARS1 in pSLA2-L of *S. rochei*. Before the discovery of the *repC* genes in pSCL2, no other genes had been reported that had homology to the *repL* genes from pSLA2-L (Hiratsu *et al.*, 2000). The research in this thesis has shown that the RepC1 and RepC2 proteins encoded by pSCL2 are very similar to the RepL1 and RepL2 proteins from pSLA2-L in both physical and structural properties. Just as RepC1 is implicated in the replication of pSCL2, RepL1 is also implicated in the replication of pSLA2-L (Hiratsu *et al.*, 2000). However, unlike the situation in pSLA2-L, only one copy of the gene encoding a RepC1 type protein was found in pSCL2. This suggests that RepC1 is essential for the replication of pSCL2. Both RepC2 and RepL2 seem to be dispensable for replication of their respective plasmids, though RepC2 may maintain a low copy number for pSCL2 while RepL2 was suggested to stabilize replication of pSLA2-L (Hiratsu *et al.*, 2000). In addition to the similarity in replication proteins, the region 606-723 bp upstream of *repC1* (Box 2, Fig. 3.6.1-3A) in pSCL2 has 79% identity to the region 577-690 bp upstream of *repL1*, which contains the putative replication origin for pSLA2-L. Furthermore, it was observed from frame analysis that two very short possible ORFs (7.207.6c and 7.207.7) located immediately upstream of the replicative genes are conserved in both pSCL2 and pSLA2-L (Fig. 3.6.1-1), though they do not show homology to any other known proteins.

The findings of this study are not the first evidence that supports the similarity in plasmid profiles between *S. rochei* and *S. clavuligerus*. Both *S. rochei* and *S. clavuligerus* have three linear plasmids, pSLA2-S (17 kb), M (100 kb), and L (206 kb) in *S. rochei*, and pSCL1 (11.7 kb), 2 (120 kb), and 3 (430 kb) in *S. clavuligerus*. Comparison of pSLA2-S and pSCL1 in previous studies demonstrated that the replication protein Rep1 from pSLA2-S has similarity to only one other protein, that encoded by ORF7 from pSCL1 (34% identity in 280 amino acids). The other essential replication protein downstream of Rep1 from pSLA2-S, Rep2, showed similarity to the protein encoded by ORF2 from pSCL1, as well as to the replicative DNA helicase DnaB proteins from many different species, including the bacteriophages lambda of *Escherichia coli*, P22 of *Salmonella typhimurium* and SPP1 of *B. subtilis* (Chang *et al.*, 1996). Furthermore, the resemblance between pSLA2-S and pSCL1 is not limited to their replication origins. Wu and Roy found the 351-bp left terminal sequence of pSCL1 had more than 70% identity to the 345-bp left terminal sequence of pSLA2-S (Wu and Roy, 1993). All of the evidence suggests that pSCL1 and pSCL2, in *S. clavuligerus*, appear to be counterparts of pSLA2-S and pSLA2-L, respectively, in *S. rochei*. Despite this similarity in plasmid profiles these two species do not show a close evolutionary relationship as judged by various genotypic and phenotypic classification approaches including numerical taxonomy, chemotaxonomy and rRNA sequence comparisons (Anderson & Wellington, 2001).

The putative replication origin of pSCL2 has been narrowed down to an approximately 2-kb sequence. In order to determine its capacity to support replication in a *Streptomyces* environment, several recombinant plasmids containing all or portions of this region were transformed into *S. lividans*. Introduction of pSG207 gave rise to higher numbers of transformants than did pSGH4. Since the insert of pSGH4 lacked 20 bp of the first hypothetical replication origin (box 1 in Fig. 3.6.1-3A), it suggests that this low GC content region might play an important role in initiation of replication of pSCL2. Previous studies of the linear plasmid SCP1 of *S. coelicolor* showed that the replication origin of SCP1 was also located in a low GC content region and contained four groups of repeats that were suggested to be important as recognition sites for replication proteins (Redenbach *et al.*, 1999). However similar long direct or inverted repeats were not found in Box 1, but rather in Box 2 in pSCL2.

In previous investigations, linear plasmids have been found not only in *Streptomyces* species, but also in other actinomycetes, such as *Planobispora*, *Rhodococcus* and *Mycobacterium* (Polo *et al.*, 1998; Crespi *et al.*, 1992; Kalkus *et al.*, 1993; Picardeau & Vincent, 1997). These linear plasmids usually contain an internal replication origin and all of them have an inverton-like structure. It was found that the replication regions of pCLP, a linear plasmid in *Mycobacterium smegmatis*, had features in common with some circular plasmids, including its partition and replication genes, and conserved sequences between these two genes (Picardeau *et al.*, 2000a). The successful transformation of plasmids containing the pSCL2 replication origin into *S. lividans* suggests that pSCL2 can also replicate as a circular form, as can pSCL1 (Shiffman & Cohen, 1992), pSLA2-S (Chang *et al.*, 1996) and pSLA2-L (Hiratsu *et al.*, 2000). However, none of these linear plasmids from *Streptomyces* show structural features similar to the replication origins of circular plasmids. The transformation experiments reported in this thesis confirm the putative replication origin in pSCL2 to be functional, but also reveal that plasmids carrying the pSCL2 replication origin have much lower transformation efficiencies than plasmids carrying an origin of replication from a circular *Streptomyces* replicon (pST1). Possibly this lower transformation efficiency is related to the fact that pST1 contains a replication origin (from SCP2*) supporting a moderate-copy-number (Larson & Hershberger, 1986) while pSCL2 supports a considerably lower copy number (estimated to be fewer than 5 copies per genome, unpublished data). Alternatively it may be that the replication origin of pSCL2, as a linear plasmid, is not able to support replication of a circular replicon as efficiently as does pST1, which contains a replication origin adapted to a circular replicon. Although replication origins in both linear and circular streptomycete plasmids use a bidirectional mechanism, other structural features in the replication origins of circular plasmids may be important for efficient replication as well.

Normally structural features in the sequence of the replication origin determine the mechanism of replication. Both linear plasmids and linear chromosomes in *Streptomyces* replicate bidirectionally, however they have different key elements for initiation of replication. The sequences of *oriC* regions in chromosomes are conserved among closely related organisms. The replication origins of various eubacteria with either linear or circular chromosomes encode a DnaA protein and contain short, conserved sequences that are essential for *oriC* function: tandem repeats of DnaA boxes, nonpalindromic 9-bp sequences and AT-rich regions (Jakimowicz *et al.*, 1998). However, there appears to be no conservation in the genetic contents and locations of ARS among the characterized linear plasmids. Besides genes encoding replication proteins, the ARS in these plasmids comprise one or more structural elements including tandem direct repeats, a series of palindromes, and low GC content sequences, but do not contain any conserved sequences. A more extensive investigation of the replication origins in linear plasmids will be necessary to elucidate the obscure features that define these regions.

Compared to the ARS, the telomeres of *Streptomyces* linear plasmids are much more conservative. The terminal sequence of pSCL2 was determined by Huang *et al.* (1998), and compared with the terminal segments of linear plasmids and linear chromosomes from other *Streptomyces* species that have no close homology. The last 180 bp of these linear molecules are 100% aligned and almost identical, containing one pair of inverted repeats and seven palindromes (the latter referring to inverted repeats with small gaps). These similarities in their sequences cause the similarities in their predicted single-stranded secondary structures, which are likely to be involved in the general terminal patching process that ensures replication of the discontinuous strand right to the end of the linear replicons (Huang *et al.*, 1998). Besides the end sequences, the terminal proteins (Tpg) and the recently discovered telomere-associated proteins (Tap) that both are essential for telomere patching are highly conserved among *Streptomyces* linear plasmids and chromosomes (Bao & Cohen, 2001; 2003; this study; Table 3.6.2-1).

4.4 What's Next?

The origin of the *Streptomyces* linear plasmids and chromosomes is a fascinating question. Chang *et al.* (1996) have suggested that the replication regions of the *Streptomyces* linear plasmids pSLA2 and pSCL1 resemble those of temperate bacteriophages of Enterobacteriaceae and *Bacillus*. It seems that linear plasmids are likely to have evolved from bacteriophages and might be responsible for the generation of linear chromosomes by integration into ancestral circular chromosomes (Volff & Altenbuchner, 2000). This integration might have happened in the chromosomal region that contains non-essential genes such as genes for secondary metabolism. It is suggested that the central core of the chromosome, which is about 5 Mb and contains all the genes essential for life, seems to be conserved during evolution (Hopwood, 2003). This hypothesis has also been proposed for other unrelated organisms. Casjens (1999) suggested that it is possible that *Borrelia* telomeres have been acquired from a poxvirus, and the *Borrelia* linear chromosomes resulted from recombination between the linear viral replicon and a circular progenitor. Similarly, the linear chromosome of *Agrobacterium tumefaciens* C58 appears to be derived from a linear plasmid (Wood *et al.*, 2001), and has been stabilized by acquisition of essential genes through successive rounds of exchange.

Determination of the DNA sequence of pSCL2 is significantly important to fully understand the biology of this linear plasmid in *S. clavuligerus*, as well as the whole class of linear plasmids in *Streptomyces* and other related species. In order to obtain the complete sequence information, two strategies can be used to fill the remaining gap regions. The first one is to produce the gap segments from PCR amplification, using the *S. clavuligerus* genomic DNA as template and the end sequences of two adjacent contigs as primers. The PCR products can be subsequently sequenced to fill the gap. This strategy has been used in this study to link some small contigs to larger ones. However, some PCR products (larger than 2 kb) were hard to sequence. An optimized sequencing technique will be helpful in resolving this problem. On the other hand, it is difficult to amplify a gap segment when the distance between the two contigs is larger than 5 kb. The alternative strategy is to screen the second library for clones that can cover these large gap regions. The clones in the second screen library contain larger inserts (3 to 6 kb). A preliminary screening has been accomplished by hybridization using the end sequences of the contigs that are close to the target gaps as probes. A further screening to more clones and using more probes needs be carried out in the future.

The ultimate purpose of the determination of the sequence of pSCL2 is to interpret the function of this linear plasmid. Up to now, the knowledge about the phenotype of pSCL2 is still quite limited. It has not been possible to isolate strains of *S. clavuligerus* cured of pSCL2, which implies that pSCL2 might be essential for *S.*

clavuligerus. Based on the current annotation, the putative *korSA* gene could play an essential role in pSCL2. *kor* (*kil* over-ride) gene is a regulator to over ride *kil* gene that facilitates horizontal transfer of circular or linear plasmids (thus also named as *tra* gene in some cases). However, expression of *kil/tra* is usually lethal to the host (*kil* stands for kill), and the presence of *kor* gene is necessary for host viability. If it is so, deletion of *korSA* in pSCL2 would cause death or poor growth of the host cells. Conjugal transfer is an important function of bacterial plasmids. The discovery of six putative transfer proteins and their potential regulators in pSCL2 provides a direction for a thorough study of the transfer-associated features of this linear plasmid in the future.

On the other hand, the finding of abundant "non-essential" and unknown genes offers a resource of enormous potential value. The sequence data and gene analysis results obtained in this study has generated an inventory of the genes in pSCL2, which may provide important information for further biological tests to investigate this linear plasmid.

BIBLIOGRAPHY

Agarwal RK and Perl A. (1993) PCR amplification of highly GC-rich DNA template after denaturation by NaOH. *Nucleic Acids Res* 21(22): 5283-4.

Aguilar A, and Hopwood DA. (1982) Determination of methylenomycin A synthesis by the pSV1 plasmid from *Streptomyces violaceusruber* SANK 95570. *J Gen Microbiol.* 128 (Pt 8): 1893-901.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol.* 215(3): 403-10.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389-402.

Anderson AS, and Wellington EM. The taxonomy of *Streptomyces* and related genera. (2001) *Int J Syst Evol Microbiol* 51(Pt 3): 797-814.

Baltz RH. Genetic manipulation of antibiotic-producing *Streptomyces*. (1998) *Trends Microbiol.* 6(2): 76-83.

Bao K, and Cohen SN. (2001) Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*. *Genes Dev* 15(12): 1518-27.

Bao K, and Cohen SN. (2003) Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. *Genes Dev* 17(6): 774-85.

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, and Hopwood DA. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417(6885): 141-7.

Bibb MJ, Findlay PR, and Johnson MW. (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* 30(1-3): 157-66.

Blackburn EH. (2001) Switching and signaling at the telomere. *Cell* 106: 661–673.

Blanco L, and Salas M. (1984) Characterization and purification of a phage ϕ29-encoded DNA polymerase required for the initiation of replication. *Proc. Natl. Acad. Sci. USA* 81: 5325–5329.

Blatch GL, and Lassle M. (1999) The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays* 21: 932-9.

Bork P, Sander C, and Valencia A. (1992) An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc Natl Acad Sci USA* 89: 7290-7294.

Brennan RG. (1993) The winged-helix DNA-binding motif: another helix-turn-helix takeoff. *Cell.* 74(5): 773-6.

Calcutt MJ. (1994) Gene organization in the dnaA-gyrA region of the *Streptomyces coelicolor* chromosome. *Gene* 151(1-2): 23-8.

Caldentey J, Blanco L, Bamford DH, and Salas M. (1993) *In vitro* replication of bacteriophage PRD1 DNA. Characterization of the protein primed initiation site. *Nucleic Acids Res.* 21: 3725–3730.

Casjens S. (1999) Evolution of the linear DNA replicons of the *Borrelia* spirochetes. *Curr. Opin. Microbiol.* 2: 529–534.

Chaconas G, Stewart PE, Tilly K, Bono JL, and Rosa P. (2001) Telomere resolution in the Lyme disease spirochete. *EMBO J.* 20(12): 3229-37.

Chang PC, and Cohen SN. (1994) Bidirectional replication from an internal origin in a linear *Streptomyces* plasmid. *Science* 265(5174): 952-4

Chang PC, Kim ES, and Cohen SN (1996) *Streptomyces* linear plasmids that contains a phage-like, centrally located, replication origin. *Mol Microbiol* 22: 789-800.

Chater KF. (1986) *Streptomyces* phage and their applications to *Streptomyces* genetics. In: Antibiotic-producing *Streptomyces*, The bacteria, Vol. IX. Edited by Queener SW and Day E. pp. 119-158 Academic Press, NY.

Chater KF, Bruton CJ, Foster SG, and Tobek I. (1985) Physical and genetic analysis of *IS110*, a transposable element of *Streptomyces coelicolor* A3(2). *Mol Gen Genet.* 200(2): 235-9.

Chater KF, Bruton CJ, Plaskitt KA, Buttner MJ, Mendez C, and Helmann JD. (1989) The developmental fate of *S. coelicolor* hyphae depends upon a gene product homologous with the motility sigma factor of *B. subtilis*. *Cell.* 59(1): 133-43.

Chen CW. (1996) Complications and implications of linear bacterial chromosomes. *Trends Genet.* 12(5): 192-6.

Chen CW, Huang CH, Lee HH, Tsai HH, and Kirby R. (2002) Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet.* 18(10): 522-9.

Chen CW, Yu TW, Lin YS, Kieser HM, and Hopwood DA. (1993) The conjugative plasmid SLP2 of *Streptomyces lividans* is a 50 kb linear molecule. *Mol Microbiol.* 7(6): 925-32.

Cheng S, Fockler C, Barnes WM, and Higuchi R. (1994) Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc Natl Acad Sci USA* 91(12): 5695-9.

Chu G, Vollrath D, and Davis RW. (1986) Separation of large DNA molecules by contour-clamped homogeneous electric fields. *Science* 234(4783): 1582-5.

Chung ST. (1987) Tn4556, a 6.8-kilobase-pair transposable element of *Streptomyces fradiae*. *J Bacteriol.* 169(10): 4436-41.

Crespi M, Messens E, Caplan AB, van Montagu M, and Desomer J. (1992) Fasciation induction by the phytopathogen *Rhodococcus fascians* depends upon a linear plasmid encoding a cytokinin synthase gene. *EMBO J* 11(3): 795-804.

Dam M, and Gerdes K. (1994) Partitioning of plasmid R1. Ten direct repeats flanking the par promoter constitute a centromere-like partition site *parC*, that expresses incompatibility. *J Mol Biol* 236: 1289-1298.

Davis MA, and Austin SJ. (1988) Recognition of the P1 plasmid centromere analog involves binding of the ParB protein and is modified by a specific host factor. *EMBO J* 7: 1881-1888.

Davis MA, Martin KA, and Austin SJ. (1992) Biochemical activities of the parA partition protein of the P1 plasmid. *Mol Microbiology* 6: 1141-1147.

de Jong RN, and van der Vliet PC. (1999) Mechanism of DNA replication in eukaryotic cells: Cellular host factors stimulating adenovirus DNA replication. *Gene* 236: 1–12.

del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M, and Diaz-Orejas R. (1998) Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev.* 62(2): 434-64.

Evans M, Kaczmarek FS, Stutzman-Engwall K, and Dyson P. (1994) Characterization of a *Streptomyces-lividans*-type site-specific DNA modification system in the avermectin-producer *Streptomyces avermitilis* permits investigation of two novel giant linear plasmids, pSA1 and pSA2. *Microbiology* 140 (Pt 6): 1367-71.

Fahnert B, Geuther R, Hoffmeier C, and Roth M. (2000) Improved determination method for plasmid copy number using pulsed field gel electrophoresis. *Biotechniques* 28(1): 26-28, 30.

Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, Lai HS, Lin SP, Qian Y, Jia HG, Najar FZ, Ren Q, Zhu H, Song L, White J, Yuan X, Clifton SW, Roe BA, and McLaughlin R. (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci U S A*. 98(8): 4658-63.

Fischer G, Holl AC, Volff JN, Vandewiele D, Decaris B, and Leblond P. (1998) Replication of the linear chromosomal DNA from the centrally located *oriC* of *Streptomyces ambofaciens* revealed by PFGE gene dosage analysis. *Res Microbiol* 149(3): 203-10.

Frangeul L, Nelson KE, Buchrieser C, Danchin A, Glaser P, and Kunst F. (1999) Cloning and assembly strategies in microbial genome projects. *Microbiology*. 145 (Pt 10): 2625-34.

Fraser CM *et al.* (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*. 390(6660): 580-6.

Gerdes K, and Molin S. (1986) Partitioning of plasmid R1. Structural and functional analysis of the *parA* locus. *J Mol Biol* 190: 269-279.

Gerdes K, Moller-Jensen J, and Bugge Jensen R. (2000) Plasmid and chromosome partitioning: surprises from phylogeny. *Mol Microbiol* 37(3): 455-66.

Gravius B, Glocker D, Pigac J, Pandza K, Hranueli D, and Cullum J. (1994) The 387 kb linear plasmid pPZG101 of *Streptomyces rimosus* and its interactions with the chromosome. Microbiology. 140 (Pt 9): 2271-7.

Griffiths AJF. (1995) Natural plasmids of filamentous fungi. *Microbiol Rev* 59: 673-685.

Haug I, Weissenborn A, Brolle D, Bentley S, Kieser T, and Altenbuchner J. (2003) *Streptomyces coelicolor* A3(2) plasmid SCP2*: deductions from the complete sequence. Microbiology. 149(Pt 2): 505-13.

Hayakawa T, Tanaka T, Sakaguchi K, Otake N, and Yonehara H. (1979) A linear plasmid-like DNA in *Streptomyces* spp. producing lankacidin group antibiotics. *J. Gen. Appl. Microbiol.* 25: 255-260.

Hayes F, Radnedge L, Davis MA, and Austin SJ. (1994) The homologous operons for P1 and P7 plasmid partition are autoregulated from dissimilar operator sites. *Mol Microbiol* 11: 249-260.

Henderson DJ, Brolle DF, Kieser T, Melton RE, and Hopwood DA (1990) Transposition of *IS117* (the *Streptomyces coelicolor* A3 (2) mini-circle) to and from a cloned target site and into secondary chromosomal sites. *Mol Gen Genet.* 224(1): 65-71.

Henikoff S. (1984) Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* 28: 351-359.

Henikoff S. (1987) Unidirectional digestion with exonuclease III in DNA sequence analysis. *Methods Enzymol.* 155: 156-165.

Higgens CE, and Kastner RE. (1971) *Streptomyces clavuligerus* sp. now., a β-lactam antibiotic producer. *Int J Syst Bacterio.* 21: 326-331.

Hinnebusch J, and Barbour AG. (1991) Linear plasmids of *Borrelia burgdorferi* have a telomeric structure and sequence similar to those of a eukaryotic virus. *J Bacteriol.* 173(22): 7233-9.

Hinnebusch J, and Tilly K. (1993) Linear plasmids and chromosomes in bacteria. *Mol Microbiol* 10(5): 917-22.

Hiratsu K, Mochizuki S, and Kinashi H. (2000) Cloning and analysis of the replication origin and the telomeres of the large linear plasmid pSLA2-L in *Streptomyces rochei. Mol Gen Genet.* 263(6): 1015-21.

Hirochika H, and Sakaguchi K. (1982) Analysis of linear plasmids isolated from *Streptomyces*: association of protein with the ends of the plasmid DNA. *Plasmid* 7(1): 59-65.

Hopwood DA. (2003) The *Streptomyces* genome-be prepared! *Nat Biotechnol.* 21(5): 505-6

Hopwood DA, Kieser T, Wright HM, and Bibb MJ. (1983) Plasmids, recombination and chromosome mapping in *Streptomyces lividans* 66. *J Gen Microbiol.* 129 (Pt 7): 2257-69.

Huang CH, Chen CY, Tsai HH, Chen C, Lin YS, and Chen CW. (2003) Linear plasmid SLP2 of *Streptomyces lividans* is a composite replicon. *Mol Microbiol.* 47(6): 1563-76.

Huang CH, Lin YS, Yang YL, Huang SW, and Chen CW. (1998) The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. *Mol Microbiol.* 28(5): 905-16.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, and Omura S. (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol.* 21(5): 526-31.

Iida S, Meyer J, and Arber W. (1983) Prokaryotic IS elements. In *Mobile Genetic Elements*: 159-221. Edited by Shapiro JA. New York: Academic Press.

Ishikawa J, and Hotta K. (1999) FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. *FEMS Microbiol Lett* 174: 251-3.

Jakimowicz D, Majka J, Messer W, Speck C, Fernandez M, Martin MC, Sanchez J, Schauwecker F, Keller U, Schrempf H, and Zakrzewska-Czerwinska J. (1998) Structural elements of the *Streptomyces* oriC region and their interactions with the DnaA protein. *Microbiology* 144 (Pt 5): 1281-90.

Janda L, Tichy P, Spizek J, and Petricek M. (1996) A deduced *Thermomonospora curvata* protein containing serine/threonine protein kinase and WD-repeat domains. *J Bacteriol*, 178: 1487-9.

Kalkus J, Dorrie C, Fischer D, Reh M, and Schlegel HG. (1993) The giant linear plasmid pHG207 from *Rhodococcus* sp. encoding hydrogen autotrophy: characterization of the plasmid and its termini. *J Gen Microbiol* 139 (Pt 9): 2055-65.

Keen CL, Mendelovitz S, Cohen G, Aharonowitz Y, and Roy KL. (1988) Isolation and characterization of a linear DNA plasmid from *Streptomyces clavuligerus*. *Mol Gen Genet* 212(1): 172-6.

Kelemen GH, and Buttner MJ. (1998) Initiation of aerial mycelium formation in *Streptomyces*. *Curr Opin Microbiol.* 1(6): 656-62.

Kelley LA, MacCallum RM, and Sternberg MJ. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 299(2): 499-520.

Kendall KJ, and Cohen SN. (1987) Plasmid transfer in *Streptomyces lividans*: identification of a *kil-kor* system associated with the transfer region of pIJ101. *J Bacteriol* 169(9): 4177-83.

Kendall KJ, and Cohen SN. (1988) Complete nucleotide sequence of the *Streptomyces lividans* plasmid pIJ101 and correlation of the sequence with genetic properties. *J Bacteriol.* 170(10): 4634-51.

Kieser T, Bibb MJ, Buttner MJ, Chater KF, and Hopwood DA (2000) Practical *Streptomyces* Genetics. The John Innes Foundation, Norwich, England.

Kieser T, and Hopwood DA. (1991) Genetic manipulation of *Streptomyces*: integrating vectors and gene replacement. *Methods Enzymol.* 204: 430-58.

Kieser HM, Kieser, T, and Hopwood DA (1992) A combined genetic and physical map of the *Streptomyces coelicolor* A3(2) chromosome. *J Bacteriol* 174(17): 5496-507.

Kinashi H, Mori E, Hatani A, and Nimi O. (1994) Isolation and characterization of linear plasmids from lankacidin-producing *Streptomyces* species. *J Antibiot* (Tokyo). 47(12): 1447-55.

Kinashi H, Shimaji M, and Sakai A. (1987) Giant linear plasmids in *Streptomyces* which code for antibiotic biosynthesis genes. *Nature* 328(6129): 454-6.

Kirby R, Wright LF, and Hopwood DA. (1975) Plasmid-determined antibiotic synthesis and resistance in *Streptomyces coelicolor*. *Nature*. 254(5497): 265-7.

Koonin EV. (1993) A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J Mol Biol* 229: 1165-1174.

Kornberg A, and Baker TA. (1992) *DNA replication*, 2nd ed. Freeman, New York.

Kurland CG, and Dong H. (1996) Bacterial growth inhibition by overproduction of protein. *Mol Microbiol*. 21(1): 1-4.

Lamb JR, Tugendreich S, and Hieter P. (1995) Tetratrico peptide repeat interactions: to TPR or not to TPR? *Trends Biochem Sci* 20(7): 257-9.

Larson JL, and Hershberger CL. (1986) The minimal replicon of a streptomycete plasmid produces an ultrahigh level of plasmid DNA. *Plasmid* 15(3): 199-209.

Lawlor EJ, Baylis HA, and Chater KF. (1987) Pleiotropic morphological and antibiotic deficiencies result from mutations in a gene encoding a tRNA-like product in *Streptomyces coelicolor* A3(2). *Genes Dev*. 1(10): 1305-10.

Leblond P, Demuyter P, Simonet JM, and Decaris B. (1991) Genetic instability and associated genome plasticity in *Streptomyces ambofaciens*: pulsed-field gel electrophoresis evidence for large DNA alterations in a limited genomic region. *J Bacteriol* 173(13): 4229-33.

Leblond P, Fischer G, Francou FX, Berger F, Guerineau M, and Decaris B. (1996) The unstable region of *Streptomyces ambofaciens* includes 210 kb terminal inverted repeats flanking the extremities of the linear chromosomal DNA. *Mol Microbiol*. 19(2): 261-71.

Leblond P, Redenbach M, and Cullum J. (1993) Physical map of the *Streptomyces lividans* 66 genome and comparison with that of the related strain *Streptomyces coelicolor* A3(2). *J Bacteriol* 175(11): 3422-9.

Leskiw BK, Mevarech M, Barritt LS, Jensen SE, Henderson DJ, Hopwood DA, Bruton CJ, and Chater KF. (1990) Discovery of an insertion sequence, IS116, from *Streptomyces clavuligerus* and its relatedness to other transposable elements from *actinomycetes*. *J Gen Microbiol* 136 (Pt 7): 1251-8.

Lewin B. (1987) The apparatus for DNA replication. *Gene III* John Wiley and Sons, NY, pp. 312-334.

Lezhava A, Mizukami T, Kajitani T, Kameoka D, Redenbach M, Shinkawa H, Nimi O, and Kinashi H. (1995) Physical map of the linear chromosome of *Streptomyces griseus*. *J Bacteriol* 177(22): 6492-8.

Lin YS, and Chen CW. (1997) Instability of artificially circularized chromosomes of *Streptomyces lividans*. *Mol Microbiol.* 26(4): 709-19.

Lin YS, Kieser HM, Hopwood DA, and Chen CW. (1993) The chromosomal DNA of *Streptomyces lividans* 66 is linear. *Mol Microbiol* 10(5): 923-33.

Llosa M, Bolland S and de la Cruz F. (1994) Genetic organization of the conjugal DNA processing region of the IncW plasmid R388. *J Mol Biol.* 235(2): 448-64.

Lu J, Ma W, Mao X, Qin ZJ, Jiang WH, and Jiao RS. (2002) Isolation and Characterization of Functional Replication Origin from *Streptomyces avermitilis*. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)* 34(6): 712-8.

Lydiate DJ, Ikeda H, and Hopwood DA. (1986) A 2.6 kb DNA sequence of *Streptomyces coelicolor* A3(2) which functions as a transposable element. *Mol Gen Genet.* 203(1): 79-88.

Madigan MT, Martinko JM, and Parker J. (2003) DNA replication, in Brock Biology of Microorganisms (10th edition). Pearson Education, Upper Saddle River, NJ, pp. 180-186.

Mahillon J, and Chandler M. Insertion sequences. (1998) *Microbiol Mol Biol Rev.* 62(3): 725-74.

Marck C. (1988) "DNA Strider": a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucleic Acids Res* 16(5): 1829-36.

Martín AC, Blanco L, García P, Salas M, and Méndez J. (1996) *In vitro* initiation of pneumococcal phage Cp-1 DNA replication occurs at the third 39 nucleotide of the linear template: a stepwise sliding-back mechanism. *J. Mol. Biol.* 260: 369–377.

McEachern MJ, Krauskopf A, and Blackburn EH. (2000) Telomeres and their control. *Annu. Rev. Genet.* 34: 331-358.

Meijer WJ, Horcajadas JA, Salas M. (2001) Phi29 family of phages. *Microbiol Mol Biol Rev.* 65(2): 261-87.

Meinhardt F, Kempken F, Kämper J, and Esser K. (1990) Linear plasmids among eukaryotes: fundamentals and applications. *Curr Genet* 17: 89-95.

Mochizuki S, Hiratsu K, Suwa M, Ishii T, Sugino F, Yamada K, and Kinashi H. (2003) The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism. *Mol Microbiol.* 48(6): 1501-10.

Mori H, Mori Y, Ichinose C, Niki H, Ogura T, Kato A, and Hiraga S. (1989) Purification and characterization of SopA and SopB proteins essential for F plasmid partitioning. *J Biol Chem* 264: 15535-15541.

Musialowski MS, Flett F, Scott GB, Hobbs G, Smith CP, and Oliver SG. (1994) Functional evidence that the principal DNA replication origin of the *Streptomyces coelicolor* chromosome is close to the dnaA-gyrB region. *J Bacteriol* 176(16): 5123-5.

Mytelka DS, and Chamberlin MJ. (1996) Analysis and suppression of DNA polymerase pauses associated with a trinucleotide consensus. *Nucleic Acids Res* 24(14): 2774-81.

Netolitzky DJ, Wu X, Jensen SE, and Roy KL. (1995) Giant linear plasmids of beta-lactam antibiotic producing *Streptomyces. FEMS Microbiol Lett* 131(1): 27-34.

Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H, Nakazawa H, Osonoe T, Kikuchi H, Shiba T, Sakaki Y, and Hattori M. (2001) Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A* 98(21): 12215-20.

Pandza S, Biukovic G, Paravic A, Dadbin A, Cullum J, and Hranueli D. (1998) Recombination between the linear plasmid pPZG101 and the linear chromosome of *Streptomyces rimosus* can lead to exchange of ends. *Mol Microbiol.* 28(6): 1165-76.

Pandza K, Pfalzer G, Cullum J, and Hranueli D. (1997) Physical mapping shows that the unstable oxytetracycline gene cluster of *Streptomyces rimosus* lies close to one end of the linear chromosome. *Microbiology.* 143 (Pt 5): 1493-501.

Pearson WR, and Lipman DJ. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A.* 85(8): 2444-8.

Picardeau M, Le Dantec C, and Vincent V. (2000a) Analysis of the internal replication region of a mycobacterial linear plasmid. *Microbiology* 146 (Pt 2): 305-13.

Picardeau M, Lobry JR, and Hinnebusch BJ. (2000b) Analyzing DNA strand compositional asymmetry to identify candidate replication origins of *Borrelia burgdorferi* linear and circular plasmids. *Genome Res.* 10(10): 1594-604.

Picardeau M, and Vincent V. (1997) Characterization of large linear plasmids in *mycobacteria. J Bacteriol* 179(8): 2753-6.

Piret JM, and Chater KF. (1985) Phage-mediated cloning of *bldA*, a region involved in *Streptomyces coelicolor* morphological development, and its analysis by genetic complementation. *J Bacteriol.* 163(3): 965-72.

Polo S, Guerini O, Sosio M, and Deho G. (1998) Identification of two linear plasmids in the actinomycete *Planobispora rosea. Microbiology* 144 (Pt 10): 2819-25.

Possoz C, Ribard C, Gagnat J, Pernodet JL, and Guerineau M. (2001) The integrative element pSAM2 from *Streptomyces*: kinetics and mode of conjugal transfer. *Mol Microbiol.* 42(1): 159-66.

Qin Z, and Cohen SN. (1998) Replication at the telomeres of the *Streptomyces* linear plasmid pSLA2. *Mol Microbiol* 28(5): 893-903.

Qin Z, and Cohen SN. (2000) Long palindromes formed in *Streptomyces* by nonrecombinational intra-strand annealing. *Genes Dev.* 14(14): 1789-96.

Ravin NV. (2003) Mechanisms of replication and telomere resolution of the linear plasmid prophage N15. *FEMS Microbiol Lett.* 221(1): 1-6.

Redenbach M, Bibb M, Gust B, Seitz B, and Spychaj A. (1999) The Linear Plasmid SCP1 of *Streptomyces coelicolor* A3(2) Possesses a Centrally Located Replication Origin and Showed Significant Homology to the Transposon Tn4811. *Plasmid* 42: 174-185.

Redenbach M, Kieser HM, Denapaite D, Eichner A, Cullum J, Kinashi H, and Hopwood DA. (1996). A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome. *Mol Microbiol* 21: 77-96.

Rees WA, Yager TD, Korte J, and von Hippel PH. (1993) Betaine can eliminate the base pair composition dependence of DNA melting. *Biochemistry* 32(1): 137-44.

Sakaguchi K. (1990) Invertrons, a class of structurally and functionally related genetic elements that includes linear DNA plasmids, transposable elements, and genomes of adeno-type viruses. *Microbiol Rev* 54(1): 66-74.

Salas M. (1991) Protein-priming of DNA replication. *Annu Rev Biochem* 60: 39-71.

Salas M, Freire R, Soengas MS, Esteban JA, Mendez J, Bravo A, Serrano M, Blasco MA, Lazaro JM, and Blanco L. (1995) Protein–nucleic acid interactions in bacteriophages ϕ29 DNA replication. *FEMS Microbiol Rev* 17: 73-82.

Sambrook, J, Fritsch EF, and Maniatis T (1989) Molecular Cloning: a laboratory manual, 2nd edition. Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press.

Scherzinger E, Haring V, Lurz R, and Otto S. (1991) Plasmid RSF1010 DNA replication *in vitro* promoted by purified RSF1010 RepA, RepB and RepC proteins. *Nucleic Acids Res.* 19(6): 1203-11.

Schultz J, Milpetz F, Bork P, and Ponting CP. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci* USA 95(11): 5857-64.

Schrempf H, Bujard H, Hopwood DA, and Goebel W. (1975) Isolation of covalently closed circular deoxyribonucleic acid from *Streptomyces coelicolor* A3(2). *J Bacteriol.* 121(2): 416-21.

Serrano M, Gutiérrez C, Freire R, Bravo A, Salas M, and Hermoso JM. (1994) Phage ϕ29 protein p6: a viral histone-like protein. *Biochimie* 76: 981–991.

Sezonov G, Possoz C, Friedmann A, Pernodet JL, and Guerineau M. (2000) KorSA from the *Streptomyces* integrative element pSAM2 is a central transcriptional repressor: target genes and binding sites. *J Bacteriol* 182(5): 1243-50.

Shiffman D, and Cohen SN. (1992) Reconstruction of a *Streptomyces* linear replicon from separately cloned DNA fragments: existence of a cryptic origin of circular replication within the linear plasmid. *Proc Natl Acad Sci USA* 89(13): 6129-33.

Smith TF, Gaitatzes C, Saxena K, and Neer EJ. (1999) The WD repeat: a common architecture for diverse functions. *Trends in Biochemical Sciences*, 24: 181-5.

Southern EM. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol.* 98(3): 503-17.

Stark MJR, Boyd A, Mileham AJ, and Romanos MA (1990) The plasmid-encoded killer system of *Kluyveromyces lactis*. *Yeast* 6: 1-29.

Stein DS, Kendall KJ, and Cohen SN. (1989) Identification and analysis of transcriptional regulatory signals for the *kil* and *kor* loci of *Streptomyces* plasmid pIJ101. *J Bacteriol* 171(11): 5768-75.

Stoytcheva Z, Joshi B, Spizek J, and Tichy P. (2000) WD-repeat protein encoding genes among prokaryotes of the *Streptomyces* genus. *Folia Microbiologica*, 45: 407-13.

Spatz K, Kohn H, and Redenbach M. (2002) Characterization of the *Streptomyces violaceoruber* SANK95570 plasmids pSV1 and pSV2 *FEMS Microbiol. Lett.* 213(1): 87-92.

Sun Y, Hegamyer G, and Colburn NH. (1993) PCR-direct sequencing of a GC-rich region by inclusion of 10% DMSO: application to mouse c-jun. *Biotechniques* 15 (3): 372-4.

Surtees JA, and Funnell BE. (1999) P1 ParB domain structure includes two independent multimerization domains. *J Bacteriol* 181: 5898-5908.

Suwa M, Sugino H, Sasaoka A, Mori E, Fujii S, Shinkawa H, Nimi O, and Kinashi H. (2000) Identification of two polyketide synthase gene clusters on the linear plasmid pSLA2-L in *Streptomyces rochei. Gene.* 246(1-2): 123-31.

Tomb JF *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature.* 388: 539-547.

Traktman P. (1996) Poxvirus DNA replication. In: DNA replication in Eukaryotic Cells (DePamphilis, M., Ed.), pp. 775-798. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Van der Vliet PC. (1995) Adenovirus DNA replication. *Curr Top Microbiol Immunol.* 199 (Pt 2): 1-30.

van der Voorn L, and Ploegh HL. (1992) The WD-40 repeat. *FEBS Lett*, 307: 131-4.

Volff JN, and Altenbuchner J. (1998) Genetic instability of the *Streptomyces* chromosome. *Mol Microbiol.* 27(2): 239-46.

Volff JN, and Altenbuchner J. (2000) A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett.* 186(2): 143-50.

Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF Jr, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D Sr, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, and Nester EW. (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* 294(5550): 2317-23.

Wu X, and Roy KL. (1993) Complete nucleotide sequence of a linear plasmid from *Streptomyces clavuligerus* and characterization of its RNA transcripts. *J. Bacteriol* 175: 37-52.

Yamasaki M, Miyashita K, Cullum J, and Kinashi H. (2000) A complex insertion sequence cluster at a point of interaction between the linear plasmid SCP1 and the linear chromosome of *Streptomyces coelicolor* A3(2). *J Bacteriol* 182: 3104-10

Zakrzewska-Czerwinska J, Jakimowicz D, Majka J, Messer W, and Schrempf H. (2000) Initiation of the *Streptomyces* chromosome replication. *Antonie Van Leeuwenhoek* 78(3-4): 211-21.

Zaman S, Radnedge L, Richards H, and Ward JM. (1993) Analysis of the site for second-strand initiation during replication of the *Streptomyces* plasmid pIJ101. *J Gen Microbiol.* 139 (Pt 4): 669-76.

Zotchev SB, and Schrempf H. (1994) The linear *Streptomyces* plasmid pBL1: analyses of transfer functions. *Mol Gen Genet* 242(4): 374-82.

Zotchev SB, Soldatova LI, Orekhov AV, and Schrempf H. (1992) Characterization of a linear extrachromosomal DNA element (pBL1) isolated after interspecific mating between *Streptomyces bambergiensis* and *S. lividans*. *Res Microbiol* 143(9): 839-45.

APPENDIX

List of primers used in PCR amplifications and sequencing reactions in this study.

No.[a]	Description	SEQUENCE (5'>3')	T_m[b]	Contig that primer located
1	SP6	GTG ACA CTA TAG AAT ACT CAA G	60	-[c]
2	T3	CTC AGA ATT AAC CCT CAC TAA	58	-
3	T7	TGA ATT GTA ATA CGA CTC ACT AT	60	-
4	SG441T3-530	CCG AGT ACG AGG CGC AGG TG	68	A8
5	SG152T7-559	CGC TCG CTC TGG CCG CTG AAG	72	A8
6	SGD9T7-1375	CGC CGA GTC ACG ATG GTT CT	64	H9
7	SG80T7-894	AGC AGA TCA GCG CCG GTG TG	66	H9
8	SG163T3-632	CGC GTC CAG ACC GTC CAT G	64	26
9	SG172T3-I	GGA GGT AGG CGG CGC AGG T	66	73
10	SG16T7-I	TGT GGT CGA GGA GCG TGC AGA	68	73
11	SG410T3-1570	CGA GGA CCA GGA GCA AGA TG	64	26
12	SG179T7-I	GCA CCG CAC CGA GCC GTT G	66	207
13	SG424T7-923	GCT CGC AGG ACG CCA GGA TC	68	424
14	SGA9T7-1303	TGC TCG CCA CGG TTG CTC CT	66	26
15	BC2T7-492	TGA GCT GCG ACC TGA CCA CAT G	70	-
16	SGB4T3-1556	CGG CAG GCG ATG GAC GTG AC	68	25
17	SG58T3-I	GGA CGC CAC TCA GCA AGA ATC	66	26
18	SG207T3-1036	GAG CGT GAG GTC GAT GAG CA	64	207
19	SG146T3-1007	GCT GCG TCT GCT CCT TCT C	62	207
20	SG3T7-589	GGC TGT GCG TCC GAG TCC T	64	25
21	SG16T7-II	CCG GAT GCG ACA CCT GTT G	62	73
22	SG172T3-II	TGC CTT GGC CGT GTC GAT C	62	73
23	SG441T7-1377	GAC GAG TTG CAT GGC GAC CT	64	A8
24	SGD6T7-404	CAA CGG CCA CTC CTA CGA AG	64	207
25	SG231T7-755	CGA GAC CGC CGA CTG CTG T	64	-
26	SG199T3-I	GCG TGC TGC CTA CTG TCG A	62	207
27	SG57T3-I	TGA GCG AGG AGA TCC AGT GT	62	A8
28	SG179T7-II	AGG CGA AGC AGC ACG GAT G	62	207
30	SG224T3-2375	CAG GAG GAG ATC GAG AAC ACT	64	68
31	NA[d]	GAG CAG GTG ACG TTC CAT C	60	-
32	SG187T3-1341	CGT GTC AAG CCG TCT GTG TG	64	207
34	SGH4T3-497	AGC GTG CGA CCG GCC ATC	62	207
35	SG172T3-III	GCC ACA GCA GCG CCA CGT	62	73
36	SGB3T3-1077	CGA ACG CAC GGT GGT CAT G	62	424

37	SG152T3-20	GGC GGA CGG AAC GGT GGA GGT	72	A8
38	SG16T3-825	CTG GCT GGT GGT GCT GGA CGA C	74	73
39	SG172T7-I	CGA AGG CGA TGC GTG AAC TGT C	70	73
40	SG187T7-365	GGC CTC CAC GCT GCT GCT	62	207
41	NA	GCG CCA TCG GTG ACC ACT CTC	70	-
42	SG136T3-457	TTG CCG GTT GTG CTG GGT TC	64	424
43	SG83T3-593	CGC TCC GGC TCC ACC TCA	62	73
44	SG188T7-468	TCG CGC CAG GCA TCG TCT TC	66	26
45	SG161T7-565	CCG CTG GCT GGC AAC GAT C	64	26
46	SG80T3-455	CGA GAA GAC GAC CTG GAC CA	64	H9
47	SGC4T3-1309	CGG TGG TGA GGT GGT TGA TGG	68	25
48	SG187T3-II	CGT CGG TGC GGA AGG TCT CG	68	207
49	SG429T7-2723	ATG CAC GCT TCT GGA GAG GT	62	424
50	SG136T7-563	AAC CCA GCA CAA CCG GCA AG	64	424
51	SG124T3-673	GCG TAA TTA GCG GAC TGA CCT	64	207
52	SG188T3-546	GCG AGC TGG GCG AGG AAC TG	68	26
53	SG67T7-I	GGA CTG GCG ACA TGC TGA TC	64	26
54	SGA4T7-696	ACC GGC GCA GCG AGC AGG A	66	26
55	SGA4T3-1673	CGA TGG AGG AGG AGG CAG TC	66	26
56	SG441T3-II	CGC TTC CTG CTT CGT CCG TGT	68	A8
57	SG298T7-3112	GCA GCG GGC TTC TCC TCA CA	66	A8
58	SG183T7-389	CCA CCC GGA AGG CTG AAC ATC	68	68
59	pSCL2-ter99	GCT GCG CGG GCC ACT CAC	64	-
60	SG54T3-I	CGC GAA CGC ATC ACC ATC AC	64	TIR
61	SGG11T3-I	GCG TCT GAC CGG CAA ATC GA	64	TIR
62	BC2T7-II	CGC CGC GCA GAT AGA TGA AG	64	-
63	SG169T3-463	GTG GAT GCT GTG GAA GTG GTC	64	69
64	contig207end1-H9	CGA GAC CGC CGA CTG CTG T	64	207
65	SG172T7-II	GGA AGC TGT GGA GGG AAT CT	62	73
66	SG22T7-I	CCA CCA CCT TCT CCT CGA TC	72	73
67	SG22T3-I	GCG GAC GGA GAA GGC CAT CA	66	73
68	2IVSG39TT-I	GCA CGT CGC GCA GAG GTC	62	68
69	2IVSG39T3-I	CAC CAC CGA ACC GAA GAA TC	62	68
70	SGB7T7-I	ACC GCA GCG AGC CTT GGA	60	207
71	SGB7T3-I	CTG GTG AGG ATG TCG AGT TC	62	207
72	SG139T7-I	CGA CGA CCC GCC AGT ACG	62	207
73	SG139T3-I	CCG GAG GCG TGA CAA TCG T	62	207
74	PCR207-II	ATC AGC ATC GAC GGT AAC AC	60	207
75	SG135T3-I	GGC GCG TTC ATC GGT CAG	60	424
76	SGD6T3-I	CGC TCG CAC CTG TCC GTT C	64	207
77	SG173T3-I	GCT CCT GCG CGT AGT CGT TC	66	207
78	2ISG12S-SP6-I	AGC GTG AGG TCG ATG AGC A	60	207
79	2ISG12S-SP6-II	CAC CAC GAC GAA ACC TCA G	60	207
80	SG139T7-73ext	CAG GGT GTT CTC GGT GGT GA	64	207
81	pSCL2-ter159	TCC CGG AGC CAT AGA CCA GTG	68	-
82	SG238T7-I	GGC CGC TCA TGT TCG ACC A	62	68

83	SGD9T7-1374	CCG CCG AGT CAC GAT GGT TC	66	H9
84	SG80T7-II	GTG GTC CAG GTC GTC TTC TC	64	H9
85	SG43T7-I	GCG AGC CGG ACG GAC AGA C	66	68
86	SG172T3-IV	GCC AAC CTC AAC CTT CAC AC	62	73
87	SGD9T7-I	CCG GTC GGC GTA GCG TGG T	66	H9
88	207-A8ext(139T3-II)	CGG GTC GTC GGA GCA CTT C	64	207
89	424-A8ext (135)	CGA CGA CGA CTT TGC TGA TG	62	424
90	A8-207ext (298)	CGG TAT CGA GGC CAC AAT G	60	A8
91	A8-424ext (441)	CCA GCC AGC CGC GTG TCT C	66	A8
92	SGA5T3-I	GAT CGG TCC GTA CAG TCT C	60	25
93	SG25T3-I	AGA GGT CGT CGA TCA GTA GTC	64	25
94	SG69T3-I	CCT TCA CGT ACC GCT CGA C	62	69
95	SGB3T7-I	CGG TGG GTG GTG TGT TCA G	62	424
96	SG68T7-I	GCC GCC GCG CCA TCG AGG A	68	68
97	SG99T3-I	GCAT CCA CCA CGG CGA CCA	62	68
98	SG113T7-I	GCC GCT GAC CCG CAC ACC A	66	68
99	SG48T3-I	TCG CTC GGC TTC GTC TTC TC	64	26
100	SG26T3-I	CAC GAA CGC ACA GAT GGA TTG	64	26
101	SG188T7-II	GCT GCG CGG CGA GGG TGT C	68	26
102	SG188T3-I'	GTC CGG TGG TGG CGT CGA	62	26
103	SG161T3-I	CGT GTC GCT CGG CTT CGT C	64	26
104	SG5T3-I	CCG TCG CGC CGT GGA GCT	64	207
105	SG179T3-I	GCC AGC CCA GGT CAC GAA G	64	207
106	SG429T3-I	GCG GCG ACC TCT CCA GAA G	64	424
107	SG16T3-I	GTC CAG GTC GGT CAG GTA GC	66	73
108	SG169T7-I	GCC GAA CAG CAT GGA GAT CA	62	69
109	SG134T7-I	CTG CCA CAT CCA TCG CTC TG	64	A8
110	SG134T3-I	GCG TAT TAC CAG TGA AGT GAC	62	A8
111	SG172T7-III	CTG GAC GAG CCG GTG GAG AC	68	73
112	SG184T3-I	GCT GGA GCG CGA GGT GAA G	64	69
113	SG180T7-I	GCG GCA GAC GAA GGC GAT G	64	H9
114	SG225T7-I / SG80T7-II	TGG CGG TAC GTC CTC CCT CT	66	H9
115	SG298T3-I	TCG CCT TCG TCT GCC GCT C	64	A8
116	SG441T7-II	GAG CTG GGA CGG TGA TGA C	62	429
117	SG59T3-I	CGC TGA TTC ACG GGA GGT TC	64	73
118	SG299T7-I	AAT GCG GCG TTC CTC TAT G	58	F11
119	SG41T7-I	TTC CGC CGC GTC GTC CAG A	64	26
120	SG188T3-II	CCT TCC TTC TCC ACC CAC TG	64	26
121	SG111T7-I	CCA CCG ACG CCG TAC TTC TC	66	73
122	2IISG12LT7-I	GTG TGT GGC GGC AGA TCA TC	64	207
123	A8-207ext2 (298)	CGC TCT GGA GGC CGT TGT C	64	A8
124	SG53T3-I	GCG GTG AGA GAG AGC AGC A	62	A8
125	SG164T7-I	CCG TAA CGA GGT GCT GAT C	60	A8
126	SG410T7-I	CCG ACA TCG ACT GGG TGC A	62	26
127	SG196T7-I	CGG CGG GCA CCA CGA TTG	62	26

128	SG19T3-I	CGC CGG ACT CCA GCA GAA G	64	26
129	SG46T3-I	GGC ACG CCC TGA GCT GGA	62	26
130	SGA5T7-I	TGA CCA CCA CTC CAG AAC TG	62	25
131	SG25T7-I	GTG CAT GGC TGG AGA ACT TC	62	25
132	SG25T3-Ia	CGC CGG TAG CCA CGA GTT C	64	25
133	SG57T7-I	GCG CGG CAC CTG GAG GCA	64	207
134	SG429T3-II	GCG AAG ACG GTG ACC AGG A	62	424
135	SGB3T3-II	GCG TTG TTC TGG CCG GAT G	62	424
136	SG424T3-I	TCT CCG CTC CGT GAT ACA G	60	424
137	SG45T3-I	GCA GAG CGA AGG AGA CCA C	62	424
138	SG45T3-III	GGG TGT CGA TTC CGG TCT G	62	424
139	SG45T7-I	TCG CCG ATG TGG ACG CCT C	64	424
140	SG224T3-III	CGT GTC GCG CAG CCA GGT GA	68	68
141	SG224T7-I	TTC ACC TGG CTG CGC GAC AC	66	68
142	SG43T7-Ia	CGA CAG TGC CAG GGC GTG A	64	68
143	SG13T3-I	CCG TTC GTA CAG GCA GTG	58	69
144	SG13-T7-I	GCG CAC TGC CTG TAC GAA C	62	69
145	SG224T3-I	CGT CGG TTC CCT TCG TGA TG	64	68
146	SG441-T7-II	GAG CTG GGA CGG TGA TGA C	62	A8
147	SG410T3-II	GGG TTC CAT TTC GAT ACG C	58	26
148	SG135T3-0	GGG TGT CGA TTC CGG TCT G	62	424

a. All the numbers for primers in this list start with "wwu"

b. The T_m of each primer was estimated using the following simplified formula:

T_m (°C) = (number of G and C) × 4°C + (number of A and T) × 2°C

c. "-" means that the clone does not belong to any contig

d. NA represents not applicable