Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*

Haruo Ikeda[1]*, Jun Ishikawa[2], Akiharu Hanamoto[3], Mayumi Shinose[3], Hisashi Kikuchi[4], Tadayoshi Shiba[5], Yoshiyuki Sakaki[6,7], Masahira Hattori[1,7]*, and Satoshi Ōmura[3,8]

Published online 14 April 2003; doi:10.1038/nbt820

Species of the genus *Streptomyces* are of major pharmaceutical interest because they synthesize a variety of bioactive secondary metabolites. We have determined the complete nucleotide sequence of the linear chromosome of *Streptomyces avermitilis*. *S. avermitilis* produces avermectins, a group of antiparasitic agents used in human and veterinary medicine. The genome contains 9,025,608 bases (average GC content, 70.7%) and encodes at least 7,574 potential open reading frames (ORFs). Thirty-five percent of the ORFs (2,664) constitute 721 paralogous families. Thirty gene clusters related to secondary metabolite biosynthesis were identified, corresponding to 6.6% of the genome. Comparison with *Streptomyces coelicolor* A3(2) revealed that an internal 6.5-Mb region in the *S. avermitilis* genome was highly conserved with respect to gene order and content, and contained all known essential genes but showed perfectly asymmetric structure at the *oriC* center. In contrast, the terminal regions were not conserved and preferentially contained nonessential genes.

S. avermitilis[1] is a Gram-positive bacterium in the genus *Streptomyces* (family Streptomycetaceae, class Actinobacteria). Streptomycetes are unique among soil bacteria because they form filamentous mycelia, aerial hyphae, and conidial spores during their life cycle[2]. Unlike other eubacterial genomes, the chromosomes of Streptomycetes form linear structures, and both ends, containing unique terminal-inverted repeats, bind terminal proteins[3]. Linear chromosomes, the predominant genetic elements in eukaryotes, have also been identified in *Borrelia burgdorferi*[4] and *Agrobacterium tumefaciens*[5,6], but no terminal-inverted repeats or terminal proteins have been identified in these chromosomes. Because of the diversity of their secondary metabolite production pathways, Streptomycetes are of great interest for the commercial production of a variety of antibiotics that are used in human and veterinary medicine and agriculture, as well as of antiparasitic agents, herbicides, pharmacologically active metabolites, and several enzymes important in the food industry and other industries[7]. Six complete genomic sequences of the class Actinobacteria are now available: two *Mycobacterium tuberculosis* strains, H37Rv[8] and CDC1551[9]; *Mycobacterium leprae* TN[10] and *Corynebacterium diphtheriae* (http://www.sanger.ac.uk/Projects/C_diphtheriae/), which are mammalian pathogens with circular chromosomes; *Corynebacterium glutamicum*[11]; and *S. coelicolor* A3(2)[12] (taxonomically belongs to *Streptomyces violaceoruber*). The availability of these genomes makes it possible to compare pathogen sequences and secondary metabolite producers within the Actinobacteria, which will provide valuable information for the application of these microbes in industrial fields including drug discovery. Here we present the complete sequence of the *S. avermitilis* genome and a comparative analysis that explores the genetic features of the Streptomycetes.

Results

Sequencing and gene annotation of the *S. avermitilis* genome. The genome sequence of *S. avermitilis* was obtained by whole genome shotgun sequencing[13] in combination with sequencing of additional cosmid clones. The principal features of the *S. avermitilis* chromosome and the linear plasmid SAP1 are summarized in Table 1. The linear chromosome contains 7,574 open reading frames (ORFs); we assigned a putative function to the encoded proteins for 4,563 (60.2%) of these. Of the remaining 3,011 ORFs, 2,738 (36.1%) showed similarity to ORFs encoding hypothetical proteins of unknown function annotated in other genomes, and 273 (3.6%) had no substantial similarity to data in the public databases. In contrast, when the *S. coelicolor* A3(2) genome was excluded from the analysis, 2,291 *S. avermitilis* ORFs (30.2%) had no substantial similarity to known genes. The plasmid SAP1 contains 96 ORFs, of which 33 (34.3%) encoded proteins that were assigned putative functions and 45 (46.9%) encoded proteins similar to proteins of unknown function annotated in other genomes. The remaining 18 ORFs (18.8%) had no substantial similarity to data in the public databases. The average GC content of the *S. avermitilis* chromosome was 70.7%, but some larger regions showed a consistently lower GC content (Fig. 1). For example, the six rRNA operons

[1]*Kitasato Institute for Life Sciences, Kitasato University, Kanagawa 228-8555, Japan.* [2]*National Institute of Infectious Diseases, Tokyo 162-8640, Japan.* [3]*The Kitasato Institute, Tokyo 108-8642, Japan.* [4]*National Institute of Technology and Evaluation, Tokyo 151-0066, Japan.* [5]*School of Science, Kitasato University, Kanagawa 228-8555, Japan.* [6]*Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan.* [7]*RIKEN Yokohama Institute, Genomic Sciences Center, Kanagawa 230-0045, Japan.* [8]*Kitasato Institute for Life Sciences, Kitasato University, Tokyo 108-8641, Japan.* **Corresponding authors (ikeda@ls.kitasato-u.ac.jp and hattori@genome.ls.kitasato-u.ac.jp).*

Table 1. Features of *S. avermitilis*

Linear chromosome		
Length (bp)		9,025,608
GC content (%)		70.7
ORFs	Conserved with protein function assigned	4,563
	Conserved with unknown protein function	2,738
	Nonconserved	273
	Total	7,574
	Average ORF size (bp)	1,034
	Coding (%)	86.2
RNA	rRNA(16S-23S-5S)	6
	tRNA	68 (43 species)
	tmRNA	1
Linear plasmid SAP1		
Length (bp)		94,287
GC content (%)		69.2
ORFs	Conserved with protein function assigned	33
	Conserved with unknown protein function	45
	Nonconserved	18
	Total	96
	Average ORF size (bp)	898
	Coding (%)	79.0
RNA	rRNA	0
	tRNA	0
	tmRNA	0

(16S-23S-5S rRNA) had GC content ranging from 57.75 to 58.01%. The rRNA operons are known to have similar GC content in all organisms, irrespective of average GC content, because of constraints on the composition of these functional RNA molecules. Transposons, including truncated forms, phage, and plasmid sequences, comprise 1.5% (99 putative transposase and 16 putative phage-plasmid integrase genes) of the *S. avermitilis* chromosome. Most of the mobile sequences (80 transposase and 7 integrase genes) were located in the regions near the two chromosomal ends, called subtelomeric regions (see below and Fig. 1).

Clustering of the 7,574 ORFs by the Basic Local Alignment Search Tool protein clustering program (BLASTCLUST; minimum 30% identity, minimum 80% length coverage) showed that 35% (2,664) clustered into 721 paralogous families, with membership ranging from 2 to 91 genes per family. Two large gene families were represented, one related to membrane-spanning components of the ATP binding cassette (ABC) transporters and the other related to two-component transcriptional regulator systems. We noted that *S. avermitilis* contains two *rpoA* genes (SAV440 and SAV4953) encoding the RNA polymerase alpha subunit. These two putative proteins are 96% identical (E value = 0) and show substantial homology with the RNA polymerase alpha subunit of *S. coelicolor* A3(2) (96% (E value = 0) and 100% (E value = 0), respectively), *Streptomyces granaticolor* (96% (E value = 0) and 99% (E value = 0), respectively), *M. tuberculosis* (78% (E value = 10^{-142}) and 75% (E value = 10^{-144}), respectively) and *M. leprae* (77% (E value = 10^{-140}) and 74% (E value = 10^{-141}), respectively). This suggested that both gene products may act as RNA polymerase alpha subunits in *S. avermitilis*. Fully sequenced eubacteria usually contain only one *rpoA* gene, and this may thus be the first case of two RNA polymerase alpha subunits being present in one bacterial genome. On the other hand, only one ORF each for *rpoB*, *rpoC*, and *rpoZ* was found in the *S. avermitilis* genome. We also identified 60 putative RNA polymerase sigma factors in the *S. avermitilis* chromosome, 47 of them belonging to the extracytoplasmic function (ECF) subfamily. *S. coelicolor* A3(2) also encodes 65 RNA polymerase sigma factors, of which 45 are ECF sigma factors[12]. The presence of numerous sigma factor genes may be a characteristic of the genus *Streptomyces*, as the next highest number known is 23 in *Mesorhizobium loti*[14]. The analysis of two RNA polymerase alpha subunits and numerous sigma factors is of importance for studying functional RNA polymerase complexes. All the annotated genes identified in this study are available on the authors' website (see URL in the Experimental Protocol).

Analysis of the linear genome structure. We identified a putative origin of replication (*oriC*) at position 5,287,935–5,289,024 of the chromosome. This region contained at least 19 *dnaA* box-like sequences[15], and the order of genes flanking the region is almost the same as that observed in circular bacterial chromosomes. The *oriC* of the linear chromosomes of *B. burgdorferi*[4], *A. tumefaciens*[5,6], and *S. coelicolor* A3(2)[12] is located in the middle of the chromosomes. In contrast, *oriC* on the *S. avermitilis* chromosome is shifted 776 kb away from the center and toward the right end. Although a GC-skew inversion is generally observed at the *oriC* of most bacterial chromosomes, no obvious GC-skew inversion could be detected on the *S. avermitilis* chromosome at any window size tested (Fig. 1A). Similarly, no GC-skew inversion has been observed in several other genomes, including *Deinococcus radiodurans* R1[16] and *Haemophilus influenzae* KW20[13]. The linear plasmid SAP1 showed no clear GC-skew inversion either but showed a marked bias in transcriptional direction of genes near the middle of the plasmid (Fig. 1B). In a few prokaryotes, including *B. burgdorferi* and *A. tumefaciens*, the termini of the linear replicons seem to be covalently closed by hairpin loop structures[4–6]. On the other hand, the termini of the *Streptomyces* chromosomes covalently bind proteins at the 5′ end[3]. Alignment of the terminal sequences of the *S. avermitilis* chromosome and SAP1 with other *Streptomyces* chromosome termini[17] indicated extensive homology in the first 96 nucleotides (see Supplementary Fig. 1 online). Recently, genes encoding the terminal proteins of the linear plasmids of *Streptomyces rochei* have been cloned and shown to be similar to genes of putative terminal proteins of other *Streptomyces*[18]. Homology searching using these sequences for the *S. avermitilis* genome identified two putative terminal protein genes, *tpgA1* and *tpgA2*, near the right end of the chromosome and the left end of SAP1, respectively. TpgA1 and TpgA2 showed 90% identity (E value = 3 10^{-92}) to each other and substantial similarity to terminal proteins of other *Streptomyces* (60–85% identity) (see Supplementary Fig. 2 online).

Comparative analysis of *S. avermitilis* with *S. coelicolor* A3(2) and other bacteria. We compared the theoretical proteome of *S. avermitilis* with those of other bacteria by using pairwise BLASTP searches without low-complexity filtering and defining reciprocal best-hit pairs as orthologs. We performed the comparative analysis with the publicly available protein sequences of *S. coelicolor* A3(2)[12], *M. tuberculosis* (AL123456), *Escherichia coli* (U00096), and *Bacillus subtilis* (AL009126). We found that 5,283 (69%) genes have orthologs in *S. coelicolor* A3(2). A smaller number, 1,966 (26%), had orthologs in *M. tuberculosis*, and only 21% had orthologs in *E. coli* (1,593 genes) or *B. subtilis* (1,586 genes). A comparison of ORFs in *S. avermitilis* and *S. coelicolor* A3(2) showed that 2,291 (738 encoding for proteins with assigned function) and 2,307 (1,080 encoding for proteins with assigned function), respectively, did not show any substantial similarity and were thus unique to each species. These unique ORFs included genes encoding for proteins involved in secondary metabolism, degradation of polymers and xenobiotics, transcriptional regulation, and transposition (Table 2).

Some differences between the phenotypes of *S. avermitilis* and *S. coelicolor* A3(2) can be explained by the presence or absence of specific orthologs in either genome. For instance, *S. avermitilis* can grow in medium containing sucrose as a sole carbon source but *S. coelicolor* A3(2) cannot. As shown in Table 2, *S. avermitilis* contains

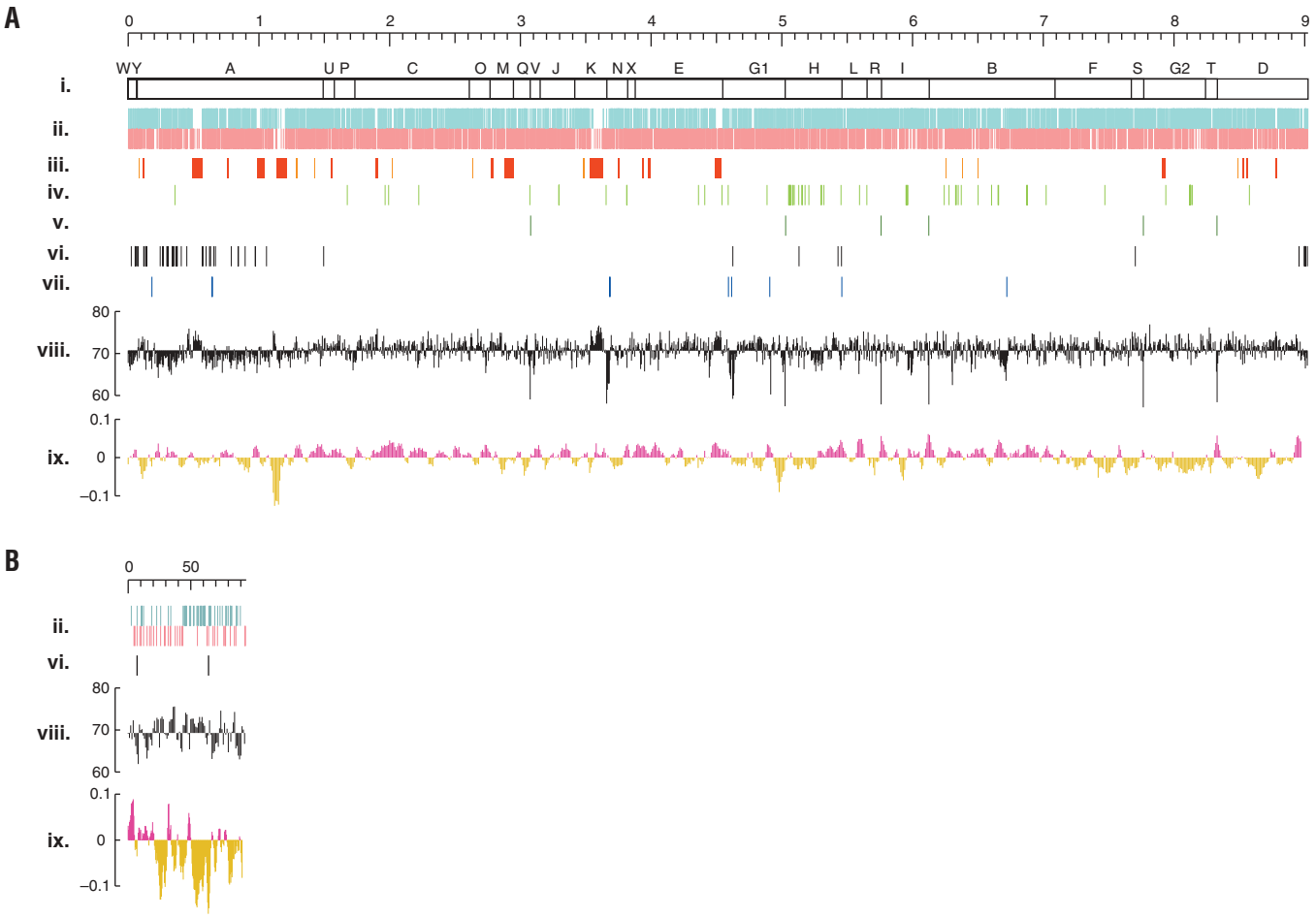

Figure 1. Schematic representation of the *S. avermitilis* chromosome (A) and the linear plasmid SAP1 (B). Each replicon is drawn to scale in megabases (A) and kilobases (B). Lines indicate *Ase*I restriction sites (i), distribution of genes according to direction of transcription (ii; + strand (upper), – strand (lower)), distribution of secondary metabolite gene clusters (iii), distribution of tRNAs (iv), distribution of rRNA operons (v), distribution of IS and transposase genes (vi), distribution of phage- or integral plasmid-related sequences (vii), GC % variation along the replicons (viii; nonoverlapping 5 kb window (A) and 1 kb window (B)), and GC skew (ix; 50 kb window and 10 kb step (A) and 5 kb window and 500 bp step (B)). The nucleotide sequence data reported in this paper appears in the DDBJ/EMBL/GenBank (NCBI) nucleotide sequence databases with the accession nos. AP005021–AP005050, BA000030, and AP005645.

a gene for a sucrose phosphotransferase enzyme (SAV3925) but *S. coelicolor* A3(2) lacks the ortholog. The presence of genes encoding enzymes involved in agar metabolism in *S. coelicolor* A3(2) (SCO3471, 5848, 5849) but not in *S. avermitilis* may explain the ability of the former to degrade agar and the lack of this capability in the latter. Both *S. avermitilis* and *S. coelicolor* A3(2)[19] lack the *recBC* genes and their corresponding suppressors *sbcAB*. These genes are involved in recombination in *E. coli*. Alternative *recD* genes, SAV5329 and SCO2737, were found in both strains, which suggests that the traditional RecBCD pathway of homologous recombination is absent in *Streptomyces* but an alternative type of recombination pathway may be present. The *S. avermitilis* chromosome contains two putative genes (SAV2423 and SAV2442) encoding the topoisomerase IV subunits involved in the separation of circular daughter chromosomes. The corresponding genes in *S. coelicolor* A3(2) are SCO5822 and SCO5836. Both chromosomes lack a XerCD-like site-specific recombination system for resolving dimeric circular chromosomes[20], a system commonly found in organisms with circular chromosomes. *S. avermitilis* lacks the operon for nitrate reductase (*narG–J*) whereas *S. coelicolor* A3(2) contains three copies of the nitrate reductase operon. The hydrogenase operons *hypA–F* and *hydAB*, which are known to be involved in nickel metabolism[21] and in urease activity[22] but have not been functionally characterized in detail, were found in *S. avermitilis* but not in *S. coelicolor* A3(2). Both *S. avermitilis* and *S. coelicolor* A3(2) have two genes—SAV3417/SAV4725 and SCO4839/SCO3334, respectively—encoding a tryptophanyl tRNA synthethase. SAV3417 showed significant similarity with the tryptophanyl tRNA synthetases of other microorganisms but only 48% identity (E value = 7 10^{-82}) with that of SAV4725. Some data suggest that SAV4725 may be involved in a mechanism that renders some *Streptomyces* strains, including *S. avermitilis* and *S. coelicolor* A3(2), naturally resistant to indolmycin, an antibiotic that inhibits tryptophanyl tRNA synthase (ref. 23 and data not shown). *S. avermitilis* and *S. coelicolor* A3(2) are naturally resistant to chloramphenicol. *S. coelicolor* encodes antibiotic-transmembrane efflux proteins CmlR and CmlR2 (SCO7526 and SCO7662), which are also found in other Actinobacteria such as *Streptomyces* and *Rhodococcus*. *S. avermitilis* lacks orthologous genes for these efflux proteins but has a gene encoding the antibiotic-modifying enzyme chloramphenicol phosphotransferase (SAV877), which has been characterized in the chloramphenicol-producing *Streptomyces venezuelae* (Q56148). These data suggest that these organisms have different mechanisms for chloramphenicol resistance.

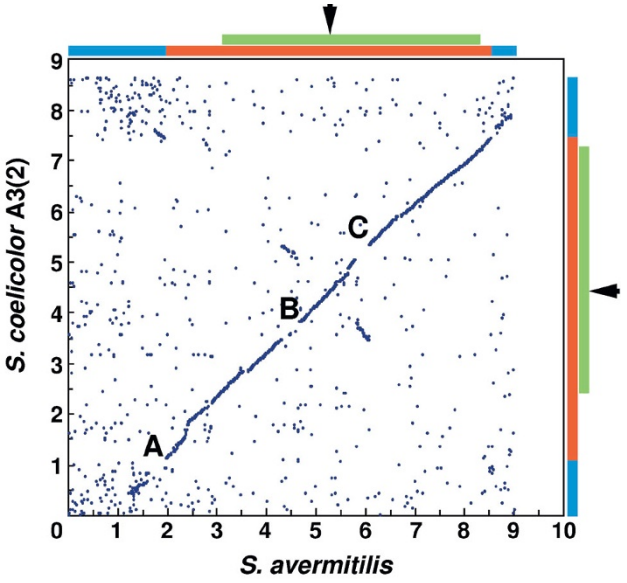

Figure 2. Synteny between *S. avermitilis* and *S. coelicolor* A3(2) linear chromosomes. Each point in this figure is a reciprocal best hit. These hits were obtained by pairwise BLASTP searches of predicted *S. avermitilis* proteins against those of *S. coelicolor* A3(2) with a maximum expectation value of 10^{-20}. Each protein pair is graphed according to the location of the corresponding gene on respective DNA molecules. The bars above and at right of plot indicate the region conserved to circular Actinobacterium chromosomes (green), subtelomeric (blue), and backbone (red) regions in *S. avermitilis* and *S. coelicolor* A3(2), respectively. Arrows indicate the position of *oriC*. A, B, and C indicate inverted regions between *S. avermitilis* and *S. coelicolor* A3(2).

The oxidation reaction carried out by cytochrome P450 using molecular oxygen often functions in the detoxification and modification of low-molecular-weight compounds. Genes encoding cytochrome P450 are not abundant in most bacteria; usually 0–4 genes are found. *S. avermitilis* contains 33 putative cytochrome P450 genes. One of them, SAV575, and four ferredoxin genes are unique to *S. avermitilis* (Table 2). *S. coelicolor* A3(2) contains 18 putative cytochrome P450 genes[24]. In *S. avermitilis*, one-third of all cytochrome P450 genes (11 genes) may be involved in the biosynthesis of secondary metabolites. The remaining 22 ORFs of *S. avermitilis* and the 18 putative cytochrome P450 genes of *S. coelicolor* A3(2) may be involved in defense mechanisms against toxic compounds in the soil environment. With 1,553 of the *S. avermitilis* and 1,227 of the *S. coelicolor* A3(2) predicted hypothetical proteins not yet characterized, much remains to be elucidated about the genetic differences between these *Streptomyces* species.

We have previously reported that *S. avermitilis* possesses at least 25 kinds of secondary metabolite gene clusters[25]. Our analysis here identified five additional secondary metabolite gene clusters in the *S. avermitilis* chromosome—four involved in the biosynthesis of terpene compounds and one in the biosynthesis of polyketide compounds. We recently confirmed experimentally the production of some metabolites predicted in this and previous studies, including geosmin, pentalenolactone, squalene, and pentaene[25] (data not shown). No secondary metabolite gene clusters were found on the plasmid SAP1. The total length of these 30 secondary metabolite gene clusters, containing 271 genes, was estimated to be 594 kb, indicating that 6.6% of the *S. avermitilis* genome is composed of genes encoding proteins involved in the biosynthesis of secondary metabolites (Fig. 1 and Supplementary Table 1 online). Many genes related to biosynthesis of secondary metabolites, including antibiotics, also seem to be less conserved among *Streptomyces* species and unique to *S. avermitilis*.

Overall genome structure of *S. avermitilis*. *S. avermitilis* and *S. coelicolor* A3(2) showed conservation of linearity and gene order along their respective chromosomes (Fig. 2). However, most of the highly conserved internal regions show a structural asymmetry between the *S. coelicolor* A3(2) and *S. avermitilis* chromosomes when the *oriC*s are placed in the same direction at the center of each chromosome. Three regions, indicated by A, B and C in Fig. 2, were also of interest. All known essential genes are located in the 6.5-Mb highly conserved internal region (SAV1625–7142 in *S. avermitilis* and SCO1196–6804 in *S. coelicolor* A3(2), respectively). Gene content and location in the 6.5-Mb conserved region also showed syntenic features with the circular chromosomes of Actinobacteria, *M. tuberculosis*, *C. diphtheriae*, and *C. glutamicum*, as shown by comparison of the *S. coelicolor* A3(2) genome with the first two of these bacteria[12]. Marked synteny was also observed in the 5.2-Mb region from SAV2398–6905 in *S. avermitilis* and the 4.9-Mb region from SCO1440–5869 in *S. coelicolor* A3(2). Region B (position 4,313,571–4,51,925; SAV3480–3709) to the left of *oriC* in the *S. avermitilis* chromosome and the corresponding region (position 5,311,553–5,063,401; SCO3062–3251) to the right of *oriC* in the *S. coelicolor* A3(2) chromosome contains essential genes including those encoding enolase, phosphoenolpyruvate carboxylase, the galactose operon (*galKET*), and a 50S ribosomal protein. Region C (position 5,834,450–6,077,662; SAV4793–5019) to the right of *oriC* in the *S. avermitilis* chromosome and the corresponding position to the left of *oriC* in the *S. coelicolor* A3(2) chromosome (position 3,781,004–3,462,666; SCO4467–4784) also contains essential genes, including those encoding two NADH dehydrogenase *nuo* operons, *rpoABC*, several 30S and 50S ribosomal proteins, and *groES–EL*.

The analysis also revealed that the regions near both telomeres are less conserved both in sequence and in regard to ortholog distribution. These variable subtelomeric regions are located 2.0 Mb from the left telomere and 0.5 Mb from the right telomere in *S. avermitilis*. Corresponding subtelomeric regions can be found 1.1 Mb from both the left and right telomeres in *S. coelicolor* A3(2). The left 2-Mb subtelomeric region contained the partially conserved and inverted A region of the *S. avermitilis* chromosome (Fig. 2) (position 1,750,041–1,957,682; SAV1418–1593) corresponding to the right subtelomeric region of the *S. coelicolor* A3(2) chromosome (position 7,602,649–7,438,541; SCO1010–1168). This inverted region contains nonessential genes such as those encoding tagatose-bisphosphate aldolase, secreted α-galactosidase, NADPH-ferredoxin reductase, and uracil DNA glycosylase. The subtelomeric regions contained 1,020 of *S. avermitilis*-specific genes corresponding to 44.5% of all 2,291 specific genes in *S. avermitilis*. Similarly, 972 (42.1%) of the 2,307 *S. coelicolor* A3(2)–specific genes were located in the subtelomeric regions.

Out of the 30 gene clusters related to secondary metabolism found in *S. avermitilis*, 17 (57%) are located in the subtelomeric regions (Fig. 1A and Supplementary Table 1 online). *S. avermitilis* specific gene clusters for the avermectin (*ave*; position 1,132,045–1,212,960) and pentaene (*pte*; position 486,648–567,017) biosynthesis pathways are also located in the left 2-Mb subtelomeric region. In contrast, many gene clusters for secondary metabolites commonly produced by several *Streptomyces* species, including *S. avermitilis*, were located in the 6.5-Mb internal conserved region of the chromosome. For example, the three biosynthetic gene clusters for geosmin (*geo*; position 2,635,583–2,640,003), pentalenolactone (*ptl*; position 3,749,307–3,758,093), and oligomycin (*olm*; position 3,536,766–3,634,592) were located in the 6.5-Mb internal region.

Discussion

We determined and analyzed the sequences of the 9.02-Mb chromosome and the plasmid SAP1 of *S. avermitilis*. The analysis of genomic sequence revealed at least 7,574 ORFs in the chromosome, of which 2,664 (35%) were shown to form 721 paralogous families, ranging from 2 to 91 genes per family. These results suggest that at least one-third of all *S. avermitilis* genes might have emerged by gene duplication during evolution. Two gene families in particular, the membrane-spanning components of the ABC transport systems and the family of two-component system transcriptional regulators, show this pattern. The paralogous families also included various genes involved in transcription, such as the RNA polymerase alpha subunit and numerous sigma factors, and genes encoding cytochrome P450, some of which may be involved in defense mechanisms against toxic compounds in the soil environment. The gene content of *S. avermitilis* suggests that its genome might have evolved by acquisition of novel gene functions to adapt to the extremely variable soil environment, the intense competition, and the drastic changes in physical conditions and nutrient availability. The abundance of such paralogous families was also observed in *S. coelicolor* A3(2) and may be characteristic of prototrophic bacteria in the genus *Streptomyces*.

Comparative analysis of *S. avermitilis* and *S. coelicolor* A3(2) revealed a 6.5-Mb, highly conserved internal region where most essential genes are located, with similar order and direction in the two species. This region also shows structural similarity to other circular bacterial chromosomes. This finding implies that the 6.5-Mb internal region is the underlying backbone of the *Streptomyces* chromosomes and may have evolved from an ancestor common to all bacteria with circular chromosomes. On the other hand, we also found variable and less conserved subtelomeric regions near both telomeres. Notably, more than half of the genes related to secondary metabolism were found in the subtelomeric regions, where no known essential gene was found. In addition, the subtelomeric regions contained most of the mobile elements in the genome and about half of the non-secondary-metabolism genes specific to *S. avermitilis*. A similar subtelomeric region organization was also found in *S. coelicolor* A3(2). These findings suggest that frequent gene duplication may have preferentially occurred in the subtelomeric regions of the *Streptomyces* chromosomes. This could also be why genes for common secondary metabolites are preferentially located in the internal region of the chromosome, whereas specific or unique genes such as *ave* are found in the subtelomeric region. The emergence of new genes and structural variability at particular loci, such as the subtelomeric regions, may be unique to linear bacterial chromosomes. The information and materials presented in this study will be of great use in improving and modifying *Streptomyces* for the production of secondary metabolites, including antibiotics.

Table 2. List of genes less conserved and unique to *S. avermitilis* or *S. coelicolor* A3(2)

SAV26, 809, 2597, 4735, 5903	Putative RNA polymerase sigma factor
SAV226	Putative reverse transcriptase homolog
SAV459, 692, 7238, 7239	Putative heat shock protein
SAV507	Putative cinnamoyl-CoA reductase
SAV563	Putative 3-(3-hydroxy-phenyl) propionate hydroxylase
SAV565	Putative 2,3-dihydroxybiphenyl-1,2-dioxygenase
SAV575	Putative cytochrome P450
SAV582, 4856, 5853, 7470	Putative ferredoxin
SAV877	Putative chloramphenicol 3-*O*-phosphotransferase
SAV1349	Putative acetoacetate decarboxylase
SAV1520	Putative 1-aminocyclopropane-1-carboxylic acid deaminase
SAV1615	Putative catechol 2,3-dioxygenase
SAV1923, 1924	Putative cytochrome D ubiquinol oxidase
SAV1941	Putative NAD-dependent formate dehydrogenase
SAV2330	Putative nitrite/sulfite reductase (*narB*)
SAV2671	Putative ATP-dependent DNA helicase (*recG*)
SAV3654	Putative 3-hydroxyacyl-ACP dehydratase (*fabA*)
SAV3925	Putative sucrose phosphotransferase enzyme
SAV7366, 7367	Putative cytochrome c_3-like hydrogenase (*hydA, B*)
SAV7373-7378	Putative [NiFe] hydrogenase (*hypA–F*)
SCO0038, 1276, 3715, 4425, 7192	Putative RNA polymerase sigma factor
SCO0107	Putative aminoglycoside nucleotidyltransferase
SCO0213	Putative nitrate/nitrite transporter protein2
SCO0216-219	Nitrate reductase operon (*narG2–J2*)
SCO0438	Pyrrolidone-carboxylate peptidase
SCO0506	NH_3-dependent NAD(+) synthetase
SCO1180	Putative DNA polymerase III beta chain
SCO2860	Rifampin ADP-ribosyl transferase
SCO3471	Extracellular agarase precursor (*dagA*)
SCO4351	Putative DNA invertase
SCO4947-4950	Nitrate reductase operon (*narG3-J3*)
SCO5848	Tagatose 6-phosphate kinase (*agaZ*)
SCO5849	AgaS protein (*agaS*)
SCO5959-5961	Cobalt transport operon (*cbiM–O*)
SCO6108	Secreted esterase (*fusH*)
SCO6532-6535	Putative nitrate reductase (*narG–J*)
SCO6824	Putative phosphonopyruvate decarboxylase
SCO7110	Ferredoxin
SCO7374	Putative oxidoreductase (*narB*)
SCO7516	Heat shock protein (*htpG*)

SAV, predicted genes of *S. avermitilis*; SCO, predicted genes of *S. coelicolor* A3(2).

Experimental protocol

Sequencing and assembly. Whole genome shotgun cloning was done using ~2 kb inserts. Shotgun template DNA was prepared by direct PCR amplification of the insert from colonies as previously reported[26] but with the following modifications to optimize amplification of the extremely high-GC DNA: dimethyl sulfoxide was added to final concentration of 5% and the denaturation temperature was 98 °C in Ex-*Taq* PCR cocktail (Takara Bio Inc., Shiga, Japan). Electrophoresis was done at reduced voltage, as this also gave better results than the standard protocol recommended by the manufacturer. Because of the high-GC homologous sequences, the first-round assembly of shotgun data using phred/phrap assembler (http://www.phrap.org/) resulted in numerous gaps and in low confidence levels for the assembled contigs. To overcome these problems, we constructed a cosmid library with ~40-kb DNA inserts. Cosmid end-sequences were used to estimate adjacent contig pairs and to evaluate correct assembly throughout the project. Cosmids were also used for gap filling by shotgun sequencing. In addition, PCR products were sequenced by primer walk to fill gaps and to resolve ambiguous regions. We also constructed *Ase*I and *Dra*I restriction maps of the *S. avermitilis* genome to confirm the final assembled data. In total, we assembled 186,619 random shotgun sequences from the whole genome, 10,787 cosmid-end sequences, 107 contig sequences from cosmid shotgun sequencing, 162 PCR-product sequences, and 72 manually curated sequences, achieving 13.3-fold coverage. The assembled data were consistent with positions and orientations of cosmid-end sequences. Both restriction sites in the assembled data were also in good agreement with the experimental data. Finally, the entire sequence was estimated to have an error rate of less than 1 per 10,000 bases (phrap score ≥ 40).

Gene finding and annotation. ORFs were predicted with Glimmer[27,28]. The program was trained with 2,000 ORFs larger than 500 bp from the genomic sequence of *S. avermitilis* as well as with the *S. coelicolor* A3(2) genes available from the public databases. We also used FramPlot[29], BLAST[30] (National Center for Biotechnology Information BLAST package; ftp://ftp.ncbi.nih.gov/), and HmmPfam[31] to confirm protein-coding genes predicted by Glimmer and found additional gene candidates.

URL. All the annotated genes identified in this study are also available on the authors' website at http://avermitilis.ls.kitasato-u.ac.jp.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgments

We thank H. Shimizu, H. Horikawa, H. Nakazawa, and T. Osonoe for technical assistance. We also thank T.D. Taylor for critical reading of this paper. This work was supported in part by Grant-in-Aid of the Ministry of Education, Culture, Sports, Science and Technology, Japan, No. 13680676, Grant of Araki Memorial Foundation for Medical and Biochemical Studies, Grant of the 21st Century COE Program, Ministry of Education, Culture, Sports, Science and Technology, Japan, and the Research for the Future Program of the Japan Society for the Promotion of Science, No. 00L01411.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 22 October 2002; accepted 22 January 2003

1. Burg, R.W. *et al.* Avermectins, new family of potent antihelminthic agents: producing organism and fermentation. *Antimicrob. Agents Chemother.* **15**, 361–367 (1979).
2. Witt, D. & Stackebrandt, E. Unification of the genera *Streptoverticillium* and *Streptomyces*, and amendation of *Streptomyces* Waksman and Henrici 1943, 339[AL]. *System. Appl. Microbiol.* **13**, 361–371 (1990).
3. Lin, Y.-S., Kieser, H.M., Hopwood, D.A. & Chen, C.W. The chromosomal DNA of *Streptomyces lividans* 66 is linear. *Mol. Microbiol.* **10**, 923–933 (1993).
4. Fraser, C.M. *et al.* Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586 (1997).
5. Wood, D.W. *et al.* The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**, 2317–2323 (2001).
6. Goodner, B. *et al.* Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**, 2323–2328 (2001).
7. Demain, A.L. Pharmaceutically active secondary metabolites of microorganisms. *Appl. Microbiol. Biotechnol.* **52**, 455–463 (1999).
8. Cole, S.T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
9. Fleischmann, R.D. *et al.* Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**, 5479–5490 (2002).
10. Cole, S.T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).
11. Mizoguchi, H. *et al.* Novel polynucleotides. European Patent Application, EP1108790 (2001).
12. Bentley, S.D. *et al.* Complete genome sequence of model of actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
13. Fleischmann, R.D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
14. Kaneko, T. *et al.* Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res.* **7**, 331–338 (2000).
15. Jakimowicz, D.M. *et al.* Structural elements of the *Streptomyces* oriC region and their interactions with the DnaA protein. *Microbiol.* **144**, 1281–1290 (1998).
16. White, O. *et al.* Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571–1577 (1999).
17. Huang, C.-H., Lin, Y.-S., Yang, Y.-L., Huang, S.-W. & Chen, C.W. The telomeres of *Streptomyces* chromosome contain conserved palindromic sequences with potential to form complex secondary structures. *Mol. Microbiol.* **28**, 905–916 (1998).
18. Bao, K. & Cohen, S.N. Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*. *Genes Dev.* **15**, 1518–1527 (2001).
19. Chen, C.W., Huang, C.-H., Lee, H.-H., Tsai, H.-H. & Kirby, R. Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet.* **18**, 522–529 (2002).
20. Chalker, A.F. *et al.* Genetic characterization of gram-positive homologs of the XerCD site-specific recombinases. *J. Mol. Microb. Biotech.* **2**, 225–233 (2000).
21. Porthun, A. Bernhard, M. & Friedrich, B. Expression of a functional NAD-reducing [NiFe] hydrogenase from the gram-positive *Rhodococcus opacus* in the gram-negative *Ralstonia eutropha*. *Arch. Microbiol.* **177**, 159–166 (2002).
22. Olson, J.W., Mehta, N.S. & Mair, R.J. Requirement of nickel metabolism proteins HypA and HypB for full activity of both hydrogenase and urease in *Helicobacter pylori*. *Mol. Microbiol.* **39**, 176–182 (2001).
23. Kitabatake, M. *et al.* Indolmycin resistance of *Streptomyces coelicolor* A3(2) by induced expression of one of its two tryptophanyl-tRNA synthetases. *J. Biol. Chem.* **277**, 23882–23887 (2002).
24. Lamb, D.C. *et al.* The cytochrome P450 complement (CYPome) of *Streptomyces coelicolor* A3(2). *J. Biol. Chem.* **277**, 24000–24005 (2002).
25. Omura, S. *et al.* Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc. Natl. Acad. Sci. USA*, **98**, 12215–12220 (2001).
26. Hattori, M. *et al.* A novel method for making nested deletions and its application for sequencing of a 300 kb region of human APP locus. *Nucleic Acids Res.* **25**, 1802–1808 (1997).
27. Salzberg, S.L., Delcher, A.L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**, 544–548 (1998).
28. Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
29. Ishikawa, J. & Hotta, K. 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**, 251–253 (1999).
30. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
31. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266 (2000).