[SV-1]

Genome Sequence of an Industrial Microorganism *Streptomyces avermitilis*

Ikeda, H.[1], Ishikawa, J.[2], Hanamoto, A.[3], Shinose, M.[3], Takahashi, C.[3], Horikawa, H.[4], Nakazawa, H.[4], Osonoe, T.[4], Kikuchi, H.[4], Shiba, T.[5], Sakaki, Y.[6,7], Hattori, M.[7] and Omura, S.[3]
[1]School of Pharmaceutical Sciences, Kitasato University, [2]National Institute of Infectious Diseases, [3]The Kitasato Institute, [4]National Institute of Technology and Evaluation, [5]School of Sciences, Kitasato University, [6]Institute of Medical Sciences, University of Tokyo, and [7]Genomic Sciences Center, RIKEN

Introduction

Streptomyces is genus of Gram-positive bacteria that grows in soil, marshes and coastal marine habitats and forms filamentous mycelium like eukaryote fungi. Morphological differentiation in *Streptomyces* involves the formation of a lawn of aerial hyphae on the colony surface that stands up into the air and differentiates into chains of spores. This process, unique among Gram-positive bacteria, requires the specialized coordination of metabolism and is more complex than other Gram-positive bacteria. The most interesting property of *Streptomyces* is their ability to produce secondary metabolites including antibiotics and bioactive compound value in human and veterinary medicine, in agriculture, and unique biochemical tools. Structural diversity is observed in these secondary metabolites which encompasses not only antibacterial, antifungal, antiviral, and antitumor compounds, but also metabolites with immunosuppresant, antihypertensive, and antihypercholesterolemic properties. Thus, *Streptomyces* are rich sources of the secondary metabolites, in which common intermediates in the cell (amino acids, sugars, fatty acids, terpenes, etc) are condensed into more complex structures by defined biochemical pathways.

Characterization of chromosome ends of eight *Streptomyces* strains has revealed evidence of linear chromosomes, indicating that chromosomal linearity might be common in the streptomycetes. Most *Streptomyces* chromosomal DNA molecules are about 8 Mb long, with terminal inverted repeats and covalently bound terminal proteins supposedly at the 5' end. This is unusually large for a bacterium, compared with well-known microorganisms as *Escherichia coli* and *Bacillus subtilis*. Streptomycetes have a higher G+C content (more than 70%) than nearly all other organisms. Thus, the *Streptomyces* chromosome is unique in its structure and size.

Streptomyces avermitilis, which was isolated from a soil sample collected at Shizuoka Prefecture, Japan in 1978, produces extremely important anthelmintic compounds "Avermectins" that are used as an antiparasitic agent since 1981, and an agricultural pesticide since 1985. Thus, *S. avermitilis* is an industrially important microorganism and the genomic information of this microorganism makes it possible to be shared not only the analysis of the industrial production process but also characteristics of *Streptomyces*. The project was started at April 2000 and to sequence the 8.8 Mb linear chromosome of *S. avermitilis* is on target for by completion until this year. We especially focus on the description of secondary metabolism in this microorganism.

1. Physical map of *S. avermitilis*

At the beginning of sequencing the genome of *S. avermitilis*, the physical map of this microorganism was determined using rare-cutter restriction enzymes. We first examined three rare-cutters, *Ase*I (ATTAAT), *Dra*I (TTTAAA), and *Ssp*I (AATATT), and these digests generated 25 fragments by *Ase*I, 9 segments by *Dra*I, and more than 30 segments by *Ssp*I. We chose to use *Ase*I for the physical map of the genome since the size range of segments were from 50 kb to 1,400 kb by *Ase*I, from 30 kb to >2,000 kb by *Dra*I, and 10 kb to 700 kb by *Ssp*I. The linking clones were isolated by insertion of the *Ase*I-fragment of a streptomycin/spectinomycin resistant gene (*aad(3")*) into the 5-kb genomic library. Each linking clone was determined by Southern hybridization of *Ase*I-cut chromosome DNA with each insert of linking clone as a probe. Two hybridized bands were detected in almost all of the hybridization experiments, but some linking clones were hybridized with 11 *Ase*I fragments that corresponded to *Ase*I-B, -D, -G1, -G2, -H, -I, -J, -R, -S, -T, and -V, in which G1 and G2 fragments overlapped. This result indicated that the linking clones contained highly homologous sequences around the *Ase*I site. The sequence linking the *Ase*I site of these clones revealed that these clones contained the *rrn* operon, in which the *Ase*I site was located in a 23S rDNA region. Cosmid clones containing a *rrn* operon were selected from the cosmid library and the regions outside of the *rrn* operon were used for hybridization probes to prevent cross-hybridization with the *rrn* region. The cosmid clones containing the *rrn* operons were classified into six groups, indicating that *S. avermitilis* has six *rrn* operons in the genome. Ultimately, we determined the physical map of *S. avermitilis* using results of linking patterns of 25 *Ase*I segments and hybridization experiments using PCR-derived amplified segments corresponding to *cysD*, *recA*, *oriC*, *proA*, and *argA* loci of *S. coelicolor* (Fig. 1).

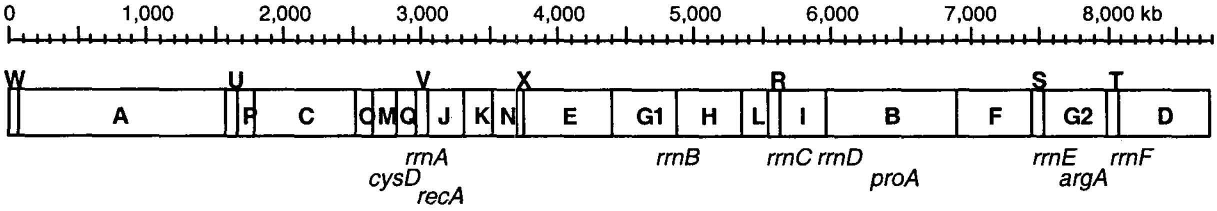

Fig. 1. Linear physical map of the chromosome*S. ofavermitilis*showing the position of known genes (*cycD*, *recA*, *proA*, and*argA*), and six*rrn* operons. Vertical lines in boxes indicate recognition sites of restriction enzyme*Ase*I. All *rrn* regions have a unique*Ase*I site between Q and V, G1and H, R and I, I and B, S and G2, and T and D.

2. Sequencing, assembly and the structure of the genome

We obtained about 200 contigs of more than 1kb by assembling all the data. The contigs could be turned into valuable 5 chains by using the linking information provided by cosmid-ends sequences. Finally, the chains were ordered and oriented on the *Ase*I physical map. The complete genome sequence is not yet determined. However, our sequence not only covers more than 99% of the genome but also gives us enough information for deducing the mechanisms of production of the secondary metabolites. Although the sequence of *S. avermitilis* genome has not been completely annotated (a few gaps remain in the assembled sequence data), we could however recognize most of the ORFs in the genome because the gaps would contain less than a few hundred bp. The total ORFs in the genome was annotated to be at least 7,800 which is about 30% and 20% more than in the genomes of *Pseudomonas aeruginosa* and the eukaryotic yeast *Saccharomyces cerecisiae*, respectively. Six *rrn* operons, in which each operon consists of the following order: 16S rDNA-23S rDNA-5S rDNA, were found. The genes coding transfer RNAs were estimated to number at least 65.

The linearity of chromosomes in the *Streptomyces* was first discovered in *S. lividans.* Other *Streptomyces* chromosomes were later found to be linear structures with two terminal inverted repeats at the both ends that were telomers. Since these terminal sequences covalently bind proteins, these protein-DNA molecules bind to glass beads and are retardated during electrophoresis. The intact chromosomal DNA preparation without the treatment of proteinase K was cut by *Ase*I and the digests were subjected to pulse-field gel electrophoresis. Two *Ase*I-fragaments, D and W, were not detected after electrophoresis, suggesting that these two fragments contain terminal ends, respectively. Two fragments of 0.9 kb and 4.5 kb were isolated from *Bam*HI-digested intact chromosome by glass-binding procedure and were hybridized to *Ase*I-W and –D, respectively. The linearity of the genome of *S. avermitilis* was also confirmed in terms of its assembly. There were no contigs at the outsides of both the *Ase*I-W and -D fragments. In the terminal regions of *Streptomyces* chromosomes examined, there is a strong homology among the first 160 nucleotides in these sequences. These first 160 nucleotide sequences were searched for homology to shotgun sequence data of *S. avermitilis* and two contigs were found to be highly homologous to the 160 nucleotides. The terminal sequence alignment of *S. avermitilis* and other four *Streptomyces* indicates that both terminal sequences of *S. avermitilis* share extensive homology to each other in the first 160 nucleotides (Fig. 2). The sizes of the terminal inverted repeats at both ends of chromosome range widely, from 24 to 550 kb in the *Streptomyces* chromosome, but the terminal inverted repeats were found in the first 174 nucleotides and long repeats such as in other *Streptomyces* chromosomes were not found in the genome of *S. avermitilis.*

```
S.lipmanii       CCCGCGGAGCGGGCCCCCCATCGCTGCGCGATGGGCA-GCGAACACCCGCGCTGCGCGCGGGTGTTGCGCTCCCGCTCCGCGGGAGCGCTGGCGGG---A
S.lividans       CCCGCGGAGCGGGTACCCTATCGCTGCGCGATAGGCAAGCGAACACCCGCGCTGCGCGCGGGTGTTGCGCTCCCGCTCCGCGGGAGCGCTGGCGGG---A
S.coelicolor     CCCGCGGAGCGGGTACCACATCGCTGCGCGATGTGCGAGCGAACACCCGGGCTGCGCCCGGGTGTTGCGCTCCCGCTCCGCGGGAGCGCTGGCGGG---A
S.avermitilis-R  CCCGCGGAGCGGGTACCACATCGCTGCGCGATGTGCAAGCGAACACCCGTGCTGCGCACGGGTGTTGCGTTCCCGCTCCGCGGGAACGCTGGCGGGGGTA
S.avermitilis-L  CCCGCGGAGCGGGTACCACATCGCTGCGCGATGTGCAAGCGAACACCCGCGCTGCGCGCGGGTGTTGCGCTCCCGCTCCGCGGGAGCGCCGGCGGG---A
pSCL2            CCCGCGGAGCGGGTACCACATCGCTTCGCGATGTGCAAGCGAACCACCGCGCTGTGCGCGGTGGTTGCGCTCCCGCTCCGCGGGAGCGCTGGAGGC---C
S.parvulus       CCCGCGGAGCGGGACCCCCACCGCTGCGCGGTGGGCAAGCGAACACCCGCGCTGCGCGCGGGTGTTGCGCTCCCGCTCCGCGTGAGCGTGAGCTGC---T
pSPA1            CCCGCGGAGCGGGTACCCCATCGCTGCGCGATGGGCAAGCGAACACCCGCGCGCAGCGCGGGTGTTGCGCTCCCGCTCCGCGTGAGCGCCCGCTGC---C
                 *************  **  * **** **** *  **  ******  *** **   ** ***  ****** ************ ** **   *  *
                 <----------->    <----------------->     <----------------------->

<---------------------> <--------

S.lipmanii       CGCTGCGC--GTCCCGCTCACCAACCCGGCTGCGCCGGGTTGGTGACGCTCCGTCCGCTGCGCTCCCGGAGCTGCGGGGCCTTCGG
S.lividans       CGCTGCGC--GTCCCGCTCACCAACCCGGCTGCGCCGGGTTGGTGACGCTCCGTCCGCTGCGCTCCCGGAGCCACGGGGCCTGCG-
S.coelicolor     CGCTGCGC--GTCCCGCTCACCAAGCCCGCTTCGCGGGCTTGGTGACGCTCCGTCCGCTGCGCTTCCGGAGTTGCGGGGCTTCGC-
S.avermitilis-R  TGCTAGGGCAGTCCCGGTCAGCAGCATTGCTGCGCAATGGTGGTGACGCTCCGTCCGCTGCGCTCCCGGAGCTGCGGTGG------
S.avermitilis-L  CGCTGCGC--GTCCCGCTCACCTCGGTTGCTGCGCAACCGGGGTGACGCTCCGTCCGCTGCGCTCCCGGAGCAGTGTGGGCTACG-
pSCL2            CGCTGCGC--GGGCCACTCACCCCCGGTGCTGTGCACCAGGGGTGAGGCTCCGTCCGCTGCGCTCCCGGAGCCATAGACCAGTGG-
S.parvulus       CGCTGCGC--GAGCAGCTCACCCGCCGCGCTTCGCACGGCGGGTGAGGCTCCGTCCGCTGCGCTCCCGGAGCAGTCGTTCCTGAC-
pSPA1            CGCTGCGC--GAGCAGCCCACCGGCCCGGCTGCGCCGGGCCGGTGACGCTCCGTCCGCTGCGCTCCCGGAGCAGCGGGGCCTACG-
                 ***  *   *  *    ** *      ***  **      ***** ***************** ******
                 ---------------->< ---------------------------> <----------------------->
```

Fig. 2. Comparison of the terminal sequences. The terminal 180 nucleotide sequences from *Streptomyces* chromosomes and linear plasmids are aligned. Palindromic sequences are indicated by the double arrows under the aligned sequences. Identities among the eight sequences are marked with the start. Both terminal sequences ofS. *avermitlis*-R and -L indicate right- and left-ends of the linear chromosome, respectively. Both pSCL2 and pSPA1 are linear plasmids ofS. *clavuligerus* and *S. parvulus*, respectively.

As shown in Fig. 3, the replication origin (*oriC*) of *S. avermitilis* was located near the middle of the linear chromosome (precisely, *oriC* was shifted from the center of the chromosome to about 600 kb toward the right end) and the replication proceeds bidirectionally towards the telomers. The gene organization within the replication origin region, where the gene order was *parB-parA-gidB-jag-orf-orf-rnpA-rpmH-dnaA-oriC-dnaN-gnd-recF-gyrB-gyrA*, is typical of bacteria possessing circular genome.

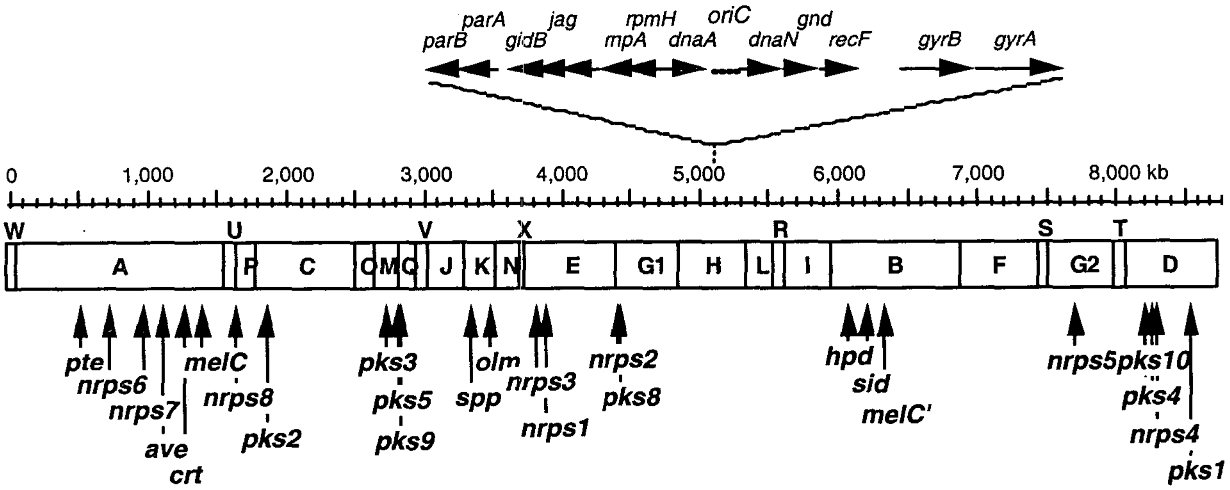

Fig. 3. The distribution of secondary metabolite clusters in the genome and region around of replication origin, and secondary metabolite clusters
Abbreviation of gene symbols*ave*, avermectin; *crt*, carotenoid; *hpd*, ochronotic pigment; *melC* and *melC'*, melanin; *nrps1-8*, peptide; *olm*, oligomycin; *pks1-10*, polyketide, *pte*, polyene macrolide; *sid*, siderophore; *spp*, spore pigment.

3. Organization of secondary metabolite clusters on the chromosome

S. avermitilis has the highest proportion of predicted secondary metabolite gene clusters of all bacterial genomes sequenced. Analysis using FramePlot, BLATP and hmmerpfam showed 25 clusters involving the biosynthesis of melanin, carotenoid, siderophore, polyketide, and peptide compounds (Fig. 3). The total lengths of these gene clusters were estimated to be about 560 kb. This analysis predicted that 6.43 % of the *S. avermitilis* genome is occupied by genes concerned with the biosyntheses of secondary metabolites, a far higher proportion than has been found in other sequenced genomes. Almost none of these secondary metabolite clusters in *S. avermitilis* were located near the center of the chromosome and more than half were in the left hand from the *oriC*. Furthermore, about half of these clusters were also found in near both ends of the chromosome. On the other hand, genes involved with primary metabolism, replication, transcription, and translation were located in a region about 6Mb from *Ase*I-C to -T fragments. These results indicate that some of secondary metabolite clusters might have been horizontally transferred from donor microorganisms in the past. Furthermore, regions near both ends contain many transposase genes, indicating that transposases played an important evolutionary role in horizontal gene transfer and also in internal genetic rearrangements in the genome. Since some transposase genes were adjacent to secondary metabolite clusters, these transposases might have been involved in the transfer of these clusters.

i) Gene clusters involving pigment and siderophore biosyntheses

S. avermitilis produces at least three kinds of melanin pigments, two are derived from tyrosine and one an aromatic polyketide. The synthesis of the former involves tyrosinase and the latter is synthesized from malonyl-CoA by a type-II PKS. These melanin pigments are produced on solid medium and then latter accumulate in the spores. Another melanin is an ochronotic pigment that is derived from homogentiginic acid and produced in both solid and liquid media. Two melanin gene clusters involving tyrosinase were found in the genome (Fig. 3), both clusters were composed of two genes, tyrosinase co-factor (MelC1) and tyrosinase (MelC2), that have been found and sequenced in seven *Streptomyces* strains. The alignment of amino acid sequences of these tyrosinases indicates that MelC2 of *S. avermitilis* is similar to that of *S. galbus*. On the other hand, MelC2' was similar to that of *S. coelicolor* which does not produce melanin. Probably, MelC2' does not function or its transcription level is too low. The genes involving melanin biosynthesis by the aromatic polyketide route have been found in most streptomycetes producing a spore pigment. The gene organization of aromatic polyketide melanin was quite similar to that in *S. coelicolor*. Another pigment gene cluster encodes the biosynthesis of a carotenoid, but the product synthesized by these genes has not yet been identified. Siderophores are involved in the transport of iron in bacteria. A gene cluster was found in the *S. avermitilis* that is presumably involved in the biosynthesis of desferrioximine derivatives because most of the genes in the cluster are quite similar to those of *Bordetella bronchiseptica* and *Sinorhizobium meliloti* responsible for desferrioxime biosynthesis.

ii) Gene clusters involving polyketide biosyntheses

Polyketides and the enzymes that make their carbon framework are ubiquitous components of microbial metabolism. *Streptomyces* and related bacteria are a rich source of structurally diverse polyketide natural products, which are derived from simple carboxylic acid precursors by a biosynthetic pathway closely analogous to the one that leads to long-chain fatty acids. There are two basic types of the PKS enzymes, iterative and modular, distinguished by both their architecture and reaction mechanism. The type-I modular PKSs consist of relatively large, multifunctional polypeptides commonly associated with the production of highly reduced metabolites such as the macrolide antibiotics. In these PKSs, each catalytic domain is used only once during assembly of the product. On the other hand, iterative PKSs consist of both fungal type-I and bacterial type–II polypeptides and each active domain is often used several times as the product is assembled. Bacterial type-II iterative PKSs are involved in the biosynthesis of aromatic polyketides.

S. avermitilis produces anthelmintic polyketide compounds, avermectins, which are the most important drugs for the treatment of endo- and ecto-parasitic infections of livestock and humans. Eight clusters containing type-I PKS genes, including the avermectin biosynthetic gene cluster, were found in the *S. avermitilis* genome (Fig. 3). The deduced amino acid sequence of each polyketide synthase was analyzed by multiple-alignment, BLASTP and hmmerpfam search programs. Fundamentally, the modular PKS contains several catalytic domains, in which the acyl-chain elongation involves □-ketoacyl-ACP synthase (KS) and acyltransferase (AT), and acyl carrier protein (ACP), and the reduction of the □-position is performed by □-ketoacyl-ACP reductase (KR), dehydratase (DH), and enoylreductase (ER).

Two of above clusters are involved in the biosyntheses of the macrocyclic lactone compounds, oligomycin and a polyene macrolide. The largest gene cluster *olm* consists of seven genes encoding a PKS carrying 17 modules including a loading module. These 17 modules contain 79 catalytic domains, but some are probably nonfunctional. On the other hand, there are five genes encoding PKS in the *pte* gene cluster (Fig. 3). These PKSs consist of 13 modules carrying 57 catalytic domains without non-functional domains. In consideration of the organization of domains in each module, five PKSs would yield a 26-membered pentaene compound.

Another five clusters were found type-I PKS genes, but the putative metabolites formed from these gene products were not identified.

In contrast to reduced polyketides assembled by modular PKS, type-II PKSs are composed of several, usually monofunctional, polypeptides, that carry out the same action repeatedly, and are involved in the synthesis of cyclic aromatic polyketides. There were three kinds of clusters containing type-II PKS genes including polyketide pigment biosyntheses as described above for melanin biosynthesis. Two of them would be involved in the synthesis of cyclic aromatic polyketides because they contain a minimal PKS unit (a mono-functional KS, chain length factor and ACP) and a DH (aromatase and cyclase having dehydration activity). Surprisingly, the cluster of pks8 consisted of two pairs of minimal PKS units. On the other hand, the cluster of pks9 had one minimal PKS unit, a KR, an aromatase and a cyclase. The phylogenetic analysis from the results of alignments of the deduced amino acid sequences of type-II ketosynthase and chain length factor indicates that the metabolite assembled by gene products of cluster of pks9 would be a decaketide because both ketosynthase and chain length factor have been classified into the group involved in the biosynthesis of decaketides.

Recently, new types of PKS gene has been reported in the genomes of *Pseudomonas* and *Streptomyces* strains, respectively. Although they have homology to plant chalcone synthase, they could not use *p*-coumaroyl-CoA as substrate and their reactions are similar to type-II PKSs that are used iteratively during chain elongation of polyketides. Pks10 has homology to these PKSs, suggesting that Pks10 is involved in the synthesis of a tetraketide or pentaketide.

iii) Gene clusters involving peptide biosyntheses

Some microorganisms contain multifunctional complexes that build specific protein templates for a ribosomal-independent biosynthesis of low molecular weight peptides of diverse structure and a broad spectrum of biological activities. Although structurally diverse, NRPS share a common mode of synthesis. Peptide bond formation takes place on a multifunctional polypeptide (NRPS) on which amino acid substrates are first activated by ATP to the corresponding adenylate. The unstable adenylate is subsequently transferred to another site of the multifunctional polypeptide where it is bound as a thioester. Thioesterified substrate amino acids are then integrated into the peptide product through a step-by-step elongation by a series of transpeptidation reactions. Thus, the synthetic reaction of NRPS is similar to that of type-I PKS.

Eight clusters containing NRPS genes were found in the *S. avermitilis* genome (Fig. 3). Although screening for peptide products synthesized by NRPSs from cultures of *S. avermitilis* has not yet been carried out, *S. avermitilis* has the ability to produce peptide products. The adenylation domain of the NRPS selects the cognate amino acid from the pool of available substrates. Recent studies have revealed that similarity between adenylation domains activating the same substrate is significantly high and there are defined general rules for the structural basis of substrate recognition by adenylation domains of NRPSs. The functional domains in each NRPS were searched for by hmmerpfam analysis. Three clusters, nrps1, nrps2 and nrps3, contain three NRPS genes, respectively. It was assumed that the peptide products synthesized by these NRPSs were tetrapeptide, hexapeptide, and dipeptide, respectively. Since the nrps6 cluster has a gene encoding a long chain fatty acid:CoA ligase and Nrps6 contains a condensation domain, the product synthesized by these gene products would be acylated. Surprisingly, the nrps7 cluster contains many genes encoding NRPSs with unusual architecture. In contrast to the common modular NRPSs consisting of multiple domains, Nrps7-2, 7-8, 7-9, 7-10, 7-12, and 7-13 are discrete polypeptides homologous to individual domains of modular NRPSs. This type of unusual NRPSs has been found in the gene cluster for bleomycin biosynthesis.

We have found 25 kinds of secondary metabolite clusters by searching for homology to polypeptides of known function involved in secondary metabolism; it thus seems that *S. avermitlis* has at least 25 secondary metabolite clusters. There are many other uncharacterized genes involving secondary metabolism in this culture. For example, the volatile substance geosmin has been detected during the cultivation of *S. avermitilis*. Why do *Streptomyces* strains produce so many kinds of secondary metabolites including antibiotics and bioactive compounds? One of the answers is that *Streptomyces* strains have many gene clusters that encode enzymes for many secondary metabolic pathways.