Finding and comparison of secondary metabolite biosynthesis gene clusters from genome sequence of Actinobacteria

(放線菌のゲノム配列からの二次代謝産物合成遺伝子クラスターの発見と比較)

千葉大学大学院医学薬学府

先端医学薬学専攻

（主任：　五ノ井　透教授）

市川　夏子

Contents

Introduction

Production of secondary metabolites is one of the industrially important features of bacteria such as actinomycetes and myxobacteria. Various secondary metabolites (or their derivatives) produced by bacteria have been developed as antibiotics, antitumor drugs and immunosuppressive drugs (1). Therefore, bacterial secondary metabolites have an important role in the development of novel medicines.

Secondary metabolites usually comprise various chemical moieties, such as polyketide backbones, amino-acid derivatives and sugars. The synthesis of secondary metabolites is catalyzed by many enzymes. Polyketides synthase (PKS) and non-ribosomal peptide synthetase (NRPS) are the major enzymes of secondary metabolite synthesis, which catalyze the elongation of polyketides and synthesis of oligopeptides, respectively. Enzymes responsible for the synthesis of other constitutive compounds, such as sugars, are often encoded by genes adjacent to PKS and NRPS genes. Through further tailoring events, such as glycosylation, alkylation and oxidation, structurally diverse and complex metabolites are finally synthesized. In addition, the production and transportation of secondary metabolites are strictly regulated by transcriptional regulators and transporters (2). Such tailoring enzymes, transcriptional regulators and transporters are also encoded by genes adjacent to PKS and NRPS genes. As a result, the whole set of genes responsible for the biosynthesis of each secondary metabolite are encoded in a large gene cluster spanning 10 kb – 100 kb.

Among many filamentous bacteria belonging to the phylum *Actinobacteria*, *Streptomyces* species have been extensively studied because of the ability to produce

various bioactive secondary metabolites. Many pioneering works about the regulation of secondary metabolism and differentiation of *Streptomyces* species were done using model organisms such as *S. coelicolor* A3(2) and *S. griseus* IFO 13350 (3, 4). The complete genome sequences of three *Streptomyces* species, *S. coelicolor* A3(2) (5), *S. avermitilis* MA-4680^{T} (6) and *S. griseus* IFO 13350 (7) have been reported and further accelerated the studies. It is revealed that many secondary metabolite biosynthesis clusters are encoded in these genomes. The abundance of biosynthesis clusters confirmed the importance of *Actinobacteria* as rich source of novel drug candidates.

In Chapter-I, we determined the complete genome sequence of *Kitasatospora setae*, and found many secondary metabolite biosynthesis clusters. It also proved the importance of the genus *Kitasatospora* as promising source of secondary metabolites. On the other hand, we also found that information related to secondary metabolite biosynthesis clusters had been poorly integrated in the process of genome annotation. In Chapter-II, we developed novel database that collects various information related to secondary metabolite biosynthesis clusters. The database can be used as useful reference to predict function of genes encoded in biosynthesis cluster. As an example of comparison analysis of biosynthesis cluster, we compared secondary metabolite biosynthesis clusters found from genus *Kitasatospora* genomes in Chapter-III. It revealed that the variety of secondary metabolite biosynthesis clusters encoded in these genomes.

CHAPTER-I

Complete genome sequence of *Kitasatospora setae* and finding of secondary metabolite biosynthesis gene clusters

1 Introduction

Kitasatospora setae is a soil-habiting mycelial bacterium belonging to the same family of *Streptomyces*, *Streptomycetaceae*. All 26 validly published species belonging to the genus *Kitasatospora* exhibit similar life style and morphology with *Streptomyces* species. *Kitasatospora* may also be comparable to *Streptomyces* in its capacity to produce bioactive secondary metabolites. The type strain of *K. setae*, NBRC 14216^{T}, is known to produce setamycin (bafilomycin B1) and bafilomycin A1, specific inhibitors of vacuolar ATPase and commonly used as biochemical reagents for investigation of molecular transport in eukaryotic cells. This genus also includes several other strains reported as producers of bioactive compounds including a proteasome inhibitor and an anti-fungal agent (8, 9).

After the first proposal of the genus *Kitasatospora* (originally *Kitasatosporia*) by Omura *et al*. in 1982 (10), the taxonomic position of *Kitasatospora* had been under a debate. It was once reclassified as a synonym of *Streptomyces* based on morphology and partial 16S rDNA analysis reported by Wellington *et al*. (11) and Ochi and Hiranuma (12). Afterward, Zhang *et al*. (13) reported that *Streptomyces* and *Kitasatospora* form distinct phyletic groups in the detailed inspection of the 16S rDNA and 16S-23S rDNA internal spacer region, and proposed the revival of the genus *Kitasatospora*. Considering the unique taxonomic position of *Kitasatospora*, as well as the common and distinct features of *Kitasatospora* and *Streptomyces*, *Kitasatospora* would be a key microorganism for further understanding of the evolution of not only actinobacteria within the family *Streptomycetaceae* but also other mycelial actinobacteria.

We determined the complete genome sequence of *K. setae* NBRC 14216^T (= KM-6054^T) and compared it with *Streptomyces* genomes. While the overall topology and gene organization of the *K. setae* NBRC 14216^T genome showed close resemblance to *Streptomyces* genomes, there are discriminative distances between them. We established robust phylogenetic position of the genus *Kitasatospora* by multi-locus phylogenetic analysis, and confirmed previous results based on 16S rDNA sequences. We also describe the possible coding capacity of the genome for bioactive secondary metabolites.

2 Materials and methods

2.1 Genome sequencing, assembly and validation

DNA shotgun libraries with average insert sizes of 1.6 kb and 6.5 kb were constructed in pUC118 (TaKaRa), while a fosmid library with average insert size of 37 kb was constructed in pCC1FOS (EPICENTRE) as described previously (14, 15). A total of 50,400 clones (34,560, 10,752 and 5,088 clones from libraries with 1.6 kb, 6.5 kb and 37 kb inserts, respectively) were subjected to sequencing from both ends of the inserts on either ABI 3730xl DNA Analyzer (Applied Biosystems) or Base Station DNA Fragment Analyzer BST-0100 (MJ Research, Inc.). Sequence reads were trimmed at a threshold quality value of 20 by Phred and assembled by Phrap and CONSED assembly tools (16, 17). For alignment and validation of contigs, Optical Mapping (OpGen) was used. Gaps between contigs were closed by sequencing PCR products which bridge two neighboring contigs. Finally, each base of *K. setae* NBRC 14216^T genome was ensured to be sequenced from multiple clones and from both directions with Phrap quality score ≧ 70 or from one direction with Phrap quality score ≧ 40. Chromosomal terminus was determined after attaching adenine and thymine homopolymers to the naked 3’ ends of

the chromosome as described previously (7).

2.2 Data analysis and annotation

The prediction of open reading frames (ORFs) was performed using Glimmer3 (18). The initial set of ORFs was manually selected from the prediction result in combination with BLASTP (19) and FramePlot (20) results. Each ORF was annotated manually using in-house genome annotation system OCSS (unpublished). Similarity search results against Uniprot (21), Interpro (22) and HAMAP (23) databases were used for functional prediction. The KEGG (24) database was used for the reconstruction of metabolic pathways. If necessary, annotation was confirmed by molecular phylogenetic analysis using ClustalW, NJplot or GARLI (http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html). Putative transporters and peptidases were independently evaluated using TransportDB (25) and MEROPS (26) databases, respectively. Non-coding genes were predicted using the Rfam (27), tRNAscan-SE (28) and ARAGORN (29) programs. Putative *oriC* region was located using originx (30) program. Putative ORFs related to mobile genetic elements were predicted and their boundaries were inferred with the assistance of GenomeMatcher (31) software. For accurate assignment of orthologs from actinobacterial genomes, comparative data compiled in the MBGD (32) database were used with further molecular phylogenetic evaluation, if necessary.

2.3 Data and strain submission

The nucleotide sequence of the *K. setae* NBRC 14216^{T} genome has been deposited in the DDBJ/EMBL/GenBank databases under accession number AP010968. The annotated genome sequence is also available at the genome database DOGAN (http://www.bio.nite.go.jp/dogan/Top). The microbial strain and genomic DNA clones used for the sequencing are available through the NBRC (NITE Biological Resource

Center; Chiba, Japan, URL:http://www.nite.go.jp/nbrc/index.html).

3 Results and discussion

3.1 General features of the *K. setae* NBRC 14216^T genome

The *K. setae* genome was composed of a single linear chromosome of 8,783,278 bp with 127,148 bp of terminal inverted repeats (TIRs). These characteristics in the genome topology were similar to those of *Streptomyces* (5–7), although the microorganism does not harbor a linear or circular plasmid. The general features of the *K. setae* genome are summarized in Table 1 and Fig. 1. The chromosome was predicted to encode 74 tRNA genes, 9 copies of ribosomal RNA operon and 7,569 protein-coding genes. Among the predicted protein-coding genes, 53.5% (4,049 ORFs) were assigned putative functions. The average G+C content of the chromosome was 74.2 %. The putative replication origin containing 20 DnaA box-like sequences was located at the center of the chromosome. This putative origin was flanked by *dnaA* and *dnaN* as in the cases of almost all bacteria, including *Streptomyces* species (33).

The finished sequence consisted of a big contig representing the major part of the chromosome connected to the same small contig at both termini. As neither sequence variation nor assembly inconsistency was found in the terminal contig, we concluded that the chromosome has inverted identical sequences at both extremities. Terminal sequences of linear chromosomes and plasmids of *Streptomyces* and *Rhodococcus* can be classified into at least 6 groups. The terminal sequence of *K. setae* chromosome was distinct from any of these groups; while the first 13 bp sequence exactly matched those of major *Streptomyces* groups I and II (Fig. S1), the subsequent region containing palindrome structures and loops, which are known in *Streptomyces* to be required for binding of telomere-associated protein (Tap), was not conserved (34, 35). Tap and

terminal protein (Tpg) encoding genes (KSE_73020 and KSE_73030) were detected in *K. setae* chromosome with lower similarity (42-45% amino acid identity) to those of *Streptomyces*. *K. setae* thus seems to have similar mechanisms for chromosome maintenance; it possesses linear chromosome with terminal inverted repeats, replicates bidirectionally from a centrally located *oriC* region, and maintains terminal sequences by Tap and Tpg. However, the binding specificity of Tap might be different even from those recognizing group I and II replicons of *Streptomyces*.

No critical difference was found between *K. setae* and *Streptomyces* species in the predicted primary metabolism such as carbohydrate metabolism, amino acid metabolism, nucleic acid metabolism and respiration. The numbers of other ubiquitous components such as membrane transporters, peptidases, transcriptional regulators and sigma factors were also almost equivalent to those of *Streptomyces*.

3.2 Taxonomic reevaluation and comparative analysis of *K. setae* NBRC 14216^{T}

From morphological similarity and rDNA relatedness to *Streptomyces* species, genus *Kitasatospora* was once regarded as a synonym of *Streptomyces* (11). Although phylogenetic analysis based on 16S rDNA sequence usually separates *Streptomyces* and *Kitasatospora* species into distinct sister groups, the results sometimes depend on the choice of the outgroup (13) or the region used for the alignment, depicting the difficulties in determining correct taxonomic relationships only from nucleotide sequences of rDNA. In order to obtain a more robust measure of the taxonomic position of *Kitasatospora* by taking advantage of its genomic information, we performed a multi-locus phylogenetic analysis using 31 conserved amino acid sequences. Ciccarelli (36) reported the construction of a tree of life across all three domains using 191 species including 14 actinobacteria. We adopted the same method by utilizing 58 actinobacterial genomes with the *E. coli* genome as an outgroup (Table S1 and S2). Amino acid sequences were aligned for each of the 31 conserved protein genes, and then all 31

alignments were concatenated and ambiguous portions were deleted before performing phylogenetic reconstruction. Fig. 2 shows phylogenetic tree obtained by Neighbor-joining method (Phylogenetic tree obtained by Maximum-likelihood method is shown in Fig. S2). Relationships between major taxonomic groups were broadly consistent with previous results obtained by 16S rDNA sequences. Within the phylogenetic tree, 82% of all predicted branches were supported by bootstrap proportions greater than 90% (i.e., 900/1000). Four *Streptomyces* species were grouped in the same branch, with *S. griseus* IFO 13350 branching out at the deepest position. *K. setae* NBRC 14216^{T} was placed within the same clan as *Streptomyces* creating the outermost branch. All these results were supported by bootstrap proportions of 100%, reinforcing the idea that the genera *Kitasatospora* and *Streptomyces* were generated from a common progenitor and have diverged into distinct sister groups. The closest to this group was *Catenulispora acidiphila* DSM 44928 (37), a mycelial actinobacterium with circular chromosome.

For further analysis of the relationship between *K. setae* and *Streptomyces*, we compared all annotated proteins of *K. setae* and four *Streptomyces* species. More than half of ORFs predicted in *K. setae* had orthologs (reciprocal best-hit pairs) in each of four *Streptomyces* species using a BLASTP threshold of $E<10^{-20}$ (Table 1). About 34% of *K. setae* ORFs had orthologs in all four *Streptomyces* genomes. The average amino acid identity between orthologous pairs from *K. setae* and *Streptomyces* species was around 60%. Despite such high similarities observed between orthologs, genome-wide comparison using ortholog plots demonstrated smaller extent of synteny between *K. setae* and *Streptomyces* genome (Fig. 3A-C). Many short synteny blocks were observed along either of the diagonal lines, suggesting that frequent inversions around the replication origin have had occurred. This is in contrast to the higher extent of colinearities with only 2-4 inversions observed in comparative analysis among *Streptomyces* genomes (5–7). Conserved core region of the *K. setae* genome predicted

by threading major synteny blocks was about 5 Mb in length ranging from KSE_13770 to KSE_57600, which is about 1 Mb smaller than the core regions deduced from comparison among *Streptomyces* genomes (7). Long range synteny between *K. setae* and *C. acidiphila* genomes was much less obvious compared with those between *K. setae* and *Streptomyces* genomes (Fig. 3D), although the number of orthologous gene pairs between *K. setae* and *C. acidiphila* (3,185) was only 10% smaller than those between *K. setae* and *Streptomyces* (3,498-3,550, see Table 1).

3.3 Gene clusters for secondary metabolite biosyntheses

Bacteria belonging to the genus *Kitasatospora* have been explored as potential new sources of various bioactive metabolites (38). *K. setae* NBRC 14216^{T} is known to produce setamycin (bafilomycin B1) which bears antitrichomonal activity (39). A total of 24 genes or gene clusters in the *K. setae* genome were predicted to be involved in the biosynthesis of secondary metabolites (Table 2). Of the 24 clusters, more than 60% (16 clusters) were located in subtelomeric regions; 11 were in the right subtelomeric region, whilst 5 were in the left subtelomeric region. The number of predicted gene clusters for secondary metabolism was slightly lower than those predicted in *Streptomyces* species (36 in *S. griseus* IFO13350, 37 in *S. avermitilis* MA-4680^{T}, 30 in *S. coelicolor* A3(2), and at least 20 in *S. scabies* 87.22), but apparently higher than those in other prokaryotes (40), underscoring the importance of the genus *Kitasatospora* as the source of bioactive compounds.

Of the 24 clusters, 5 were estimated for terpene biosynthesis, 12 for polyketides or non-ribosomal peptides, 2 for siderophores, and 5 for lantibiotics and others. Five ORFs containing terpene synthase domain (IPR005630) were classified by the phylogenetic analysis described by Komatsu *et al*.(41), indicating that KSE_46080 was in the group of germacradienol/geosmin synthase (42, 43) and KSE_70210 in the group of 2-methylisoborneol synthase. Consistently, geosmin and 2-methylisoborneol were

identified from the culture of *K. setae*. Of the 12 clusters for polyketides or non-ribosomal peptides, setamycin (bafilomycin B1) cluster was estimated and experimentally proved to be KSE_73410 to KSE_73580 (H. Ikeda, unpublished results). An 84 kbp region (KSE_70410-KSE_70650) containing the cluster KSE_70570 to KSE_70620 showed striking resemblance in features of gene organization and in deduced amino acid sequences of each ORF to the kirromycin biosynthetic gene cluster, the first characterized combined *cis*- *trans*-PKS cluster in *Streptomyces collinus* Tu 365 (44). This region was experimentally confirmed to be responsible for the biosynthesis of factumycin, an antibiotic structurally related to kirromycin (H. Ikeda, unpublished results). Regarding the 2 siderophore clusters, the one (KSE_12660 to KSE_12700) showed good similarity in gene arrangement and deduced amino sequence of each ORF to a cluster in *S. avermitilis* (SAV_7320 to SAV_7323, lacking a homolog corresponding to KSE_12680), both of which were similar to vibrioferrin cluster identified in *Vibrio parahaemolyticus* (45). Another one (KSE_53800 to KSE_53830) was similar in gene arrangement to the rhizobactin cluster in *Sinorhizobium meliloti* (46), but differed from any of the known siderophore biosynthesis clusters of *Streptomyces*. Interestingly, *K. setae* lacks a gene cluster for nocardamin (desferrioxamine) biosynthesis that was commonly found in genome-sequenced *Streptomyces* species (40, 47).

The initiation of secondary metabolite synthesis is known to be linked with morphological differentiation in *Streptomyces griseus* via the gamma-butyrolactone autoregulator cascade. In this regard, it would be noteworthy that *K. setae* possesses three homologs of the autoregulator receptor, KsbA (KSE_58650), KsbB (KSE_01050 and KSE_75690; identical genes encoded in the terminal inverted repeat) and KsbC (KSE_44580). KsbA in *K. setae* was experimentally confirmed to be involved only in secondary metabolism (48),while the involvement of other two remains to be clarified in the future.

3.4 Findings after *K. setae* genome publish

After we published the genome sequence of *K. setae*, several findings have been reported on the basis of our reported genomic features. Especially, abundance of secondary metabolite biosynthesis clusters drew scientists' attention and some biosynthesis clusters have proved to produce compound. Aroonsri *et al*. characterized the regulation mechanism of secondary metabolite production using KsbC (KSE_44580) gene (49). Deletion of the gene *ksbC* resulted in lowered production of bafilomycin and enhanced production of a novel beta-carboline alkaloid, kitasetaline. They also started a screening with a genomic BAC library of *K. setae*, and they successfully identified kitasetaline biosynthesis clusters from *K. setae* genome (KSE_70630 - KSE_70700). Setamycin (bafilomycin B1) biosynthesis cluster, we assigned in *K. setae* genome (KSE_73410-KSE_73580), was also found in *Streptomyces griseus* DSM 2608 and characterized (50).

According to the genome sequence of *K. setae,* the taxonomic status of the genus *Kitasatospora* attracted attentions and remarked to be important using genomic sequence for phylogenetic comparison. Girard *et al*. focused on the diversification of *K.setae* SsgA-like family proteins (SALPs), likely involved in the regulation of peptidoglycan synthesis in sporogenic cell division, and proposed that SsgA and SsgB proteins can be used as taxonomic marker for classification of morphologically complex actinomycetes (51, 52).

4 Conclusion

Streptomycetes are thought to have emerged about 440 million years ago. These groups of bacterium are some of the most highly differentiated microorganisms with a complex life cycle. The members of the family *Streptomycetaceae* are known to be difficult to distinguish using phenotypic properties. Here, we analyzed the first genome

sequence of a *Streptomycetaceae* bacterium other than *Streptomyces* species. Phylogenetic analysis based on amino acid sequences, together with genome-wide comparison of the predicted genes suggests that the genera *Streptomyces* and *Kitasatospora* were diverged directly from their last common ancestor. The chromosomal linearity and the presence of terminal inverted repeat sequence would also be the features inherited from the common ancestor.

Here, we report the existence of more than 24 secondary metabolite biosynthesis clusters. Apart from taxonomic distinctness with Streptomyces, the genus *Kitasatospora* is recognized as the important source of bioactive compounds. Though a majority of these gene clusters play unknown roles in the secondary metabolites biosynthesis processes, consecutive reports of novel compounds from *Kitasatospora* species will unveil their roles.

Table 1. General features of *K. setae* NBRC 14216^{T} genome and *Streptomyces* genomes

	Length (bp)	TIR (bp)	G+C Content (%)	CDS (no.)	rRNA operons (no.)	tRNA genes (no.)	Avg. CDS length (bp)	Coding dentity (%)	Reciprocal BLAST best hit pair (no.)[b]
K. setae NBRC 14216^{T}	8,783,278	127,148	74.2	7,569	9	74	1,012	87.0	-
S. coelicolor A3(2)	8,667,507	21,653	72.1	7,825	6	63	991	88.9	3,550
S. avermitilis MA-4680^{T} [a]	9,025,608	49	70.7	7,582	6	68	1,027	86.3	3,498
S. griseus IFO13350	8,545,929	132,910	72.2	7,138	6	66	1,055	88.1	3,513
S. scabies 87.22	10,148,695	18,488	71.5	8,746	6	75	1,005	86.2	3,534

[a] Based on the latest annotation data of *S. avermitilis* MA-4680^{T} maintained by H. Ikeda (http://avermitilis.ls.kitasato-u.ac.jp)

[b] Numbers of best hit pair between *K. setae* NBRC 14216^{T} and each *Streptomyces* are calculated using BLASTP program with a threshold E value of $1e^{-20}$.

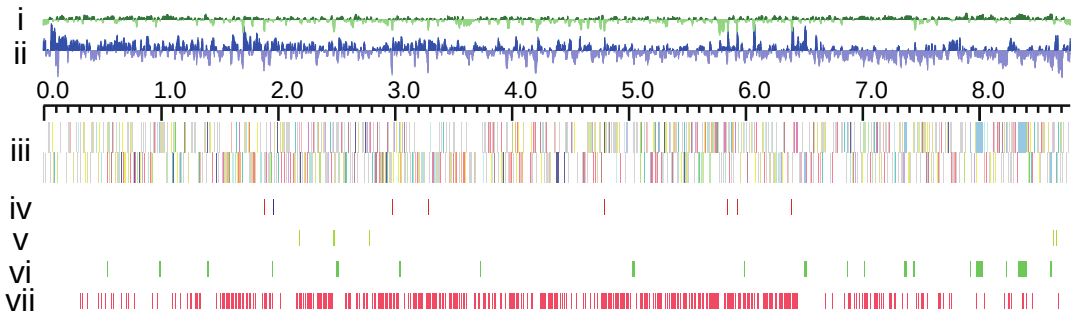


Figure 1. Schematic representation of the *K. setae* NBRC 14216^{T} chromosome. (i) G+C content; (ii) GC-skew; (iii) ORFs encoded in forward (upper) and reverse (lower) strand. Each ORF is colored on the basis of the predicted function. (iv) RNA encoding genes. rRNA operons are colored in red and tRNAs are colored in blue. (v) putative insertion sequences. (vi) secondary metabolism gene clusters. (vii) Red bars indicate ORFs conserved commonly in all four *Streptomyces* genome in the 'core region'

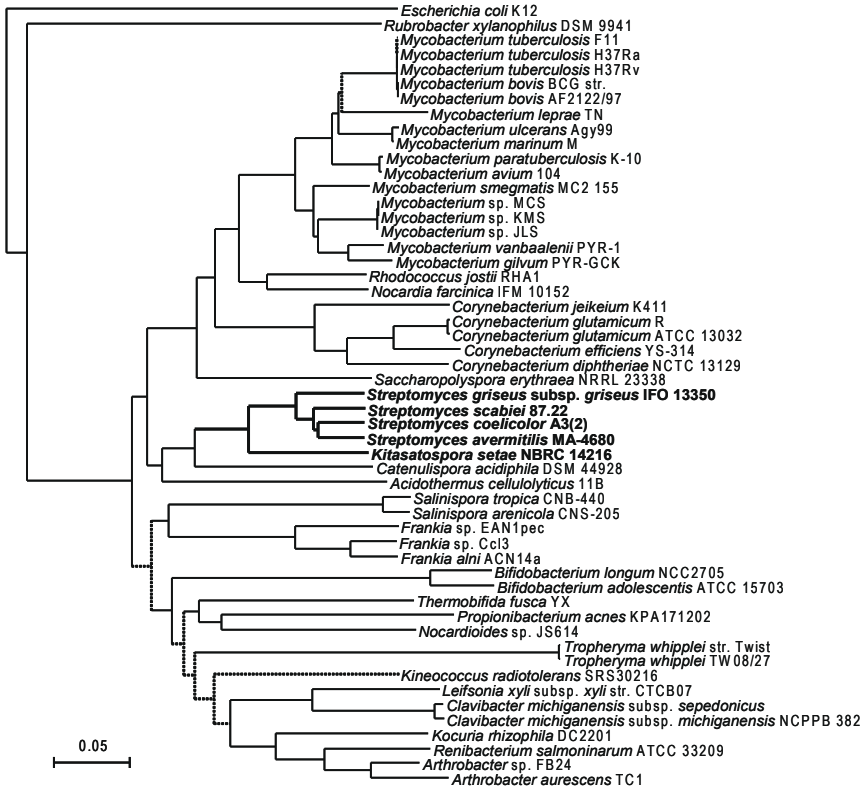


Figure 2. Phylogenetic tree based upon amino acid sequences of 31 protein-coding genes analyzed by Neighbor-joining method. Branches with less than 90% bootstrap support are represented in dashed lines. Lists of organisms and genes used for the analysis are shown in supplementary materials Table S1 and S2, respectively. Names of the organisms mentioned in the text are shown in bold type.

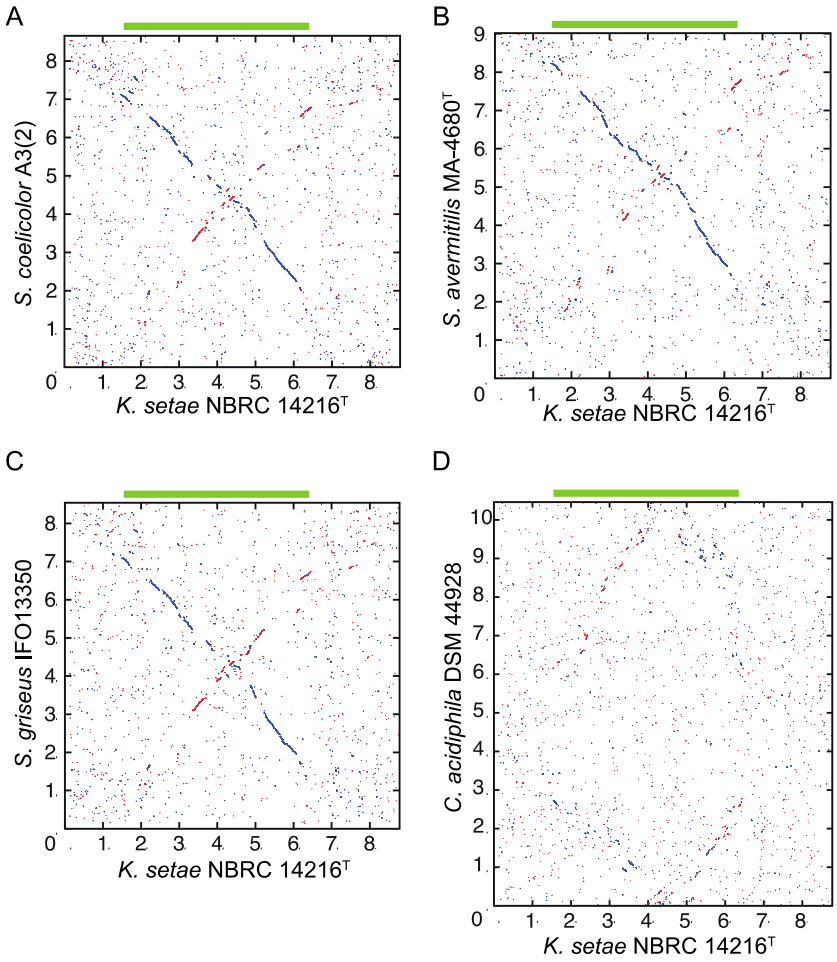


Figure 3. Synteny between the genomes of *K. setae* NBRC 14216^T and *S. coelicolor* A3(2) (A), *K. setae* NBRC 14216^T and *S. avermitilis* MA-4680^T (B), *K. setae* NBRC 14216^T and *S. griseus* IFO 13350 (C), and *K. setae* NBRC 14216^T and *C. acidiphila* DSM 44928 (D). Reciprocal BLAST best-hit pairs with a threshold value of E < 10^{-20} were plotted. The direction of each chromosome was adjusted so that the *dnaA* gene faces the same direction. Green bar in each panel represents the conserved core region on the *K. setae* chromosome.

Table 2. Biosynthetic genes of secondary metabolites in *K. setae* NBRC 14216^{T}

orfid	note
terpene	
KSE_00200 (=KSE_76540)[a]	sesquiterpene
KSE_12950	sesquiterpene
KSE_17590-KSE_17630	squalene/hopanoid
KSE_46080	germacradienol/geosmin
KSE_70210-KSE_70220	2-methylisoborneol
NRPS and PKS	
KSE_18000	type III PKS
KSE_22630-KSE_22810	discrete NRPS
KSE_27200-KSE_27290	type I PKS
KSE_33340-KSE_33360	discrete NRPS
KSE_58150-KSE_58200	NRPS
KSE_61120	NRPS (incomplete)
KSE_65510-KSE_65560	NRPS and type I PKS
KSE_65960-KSE_66030	NRPS and type II PKS
KSE_70570-KSE_70620	type I PKS and NRPS (factumycin)
KSE_72410-KSE_72480	type II PKS (spore pigment)
KSE_73410-KSE_73580	type I PKS (bafilomycins)
KSE_75420-KSE_75430	NRPS
other	
KSE_04750-KSE_04770	lantibiotic
KSE_09030-KSE_09170	similar to valanimycin biosynthetic genes
KSE_12660-KSE_12700	siderophore
KSE_27300-KSE_27440	similar to valanimycin biosynthetic genes
KSE_45610-KSE_45680	lantibiotic
KSE_58810-KSE_58830	lantibiotic
KSE_53800-KSE_53830	siderophore

[a] KSE_00200 was embedded in terminal inverted repeat and identical to KSE_76540.

Fig. S1

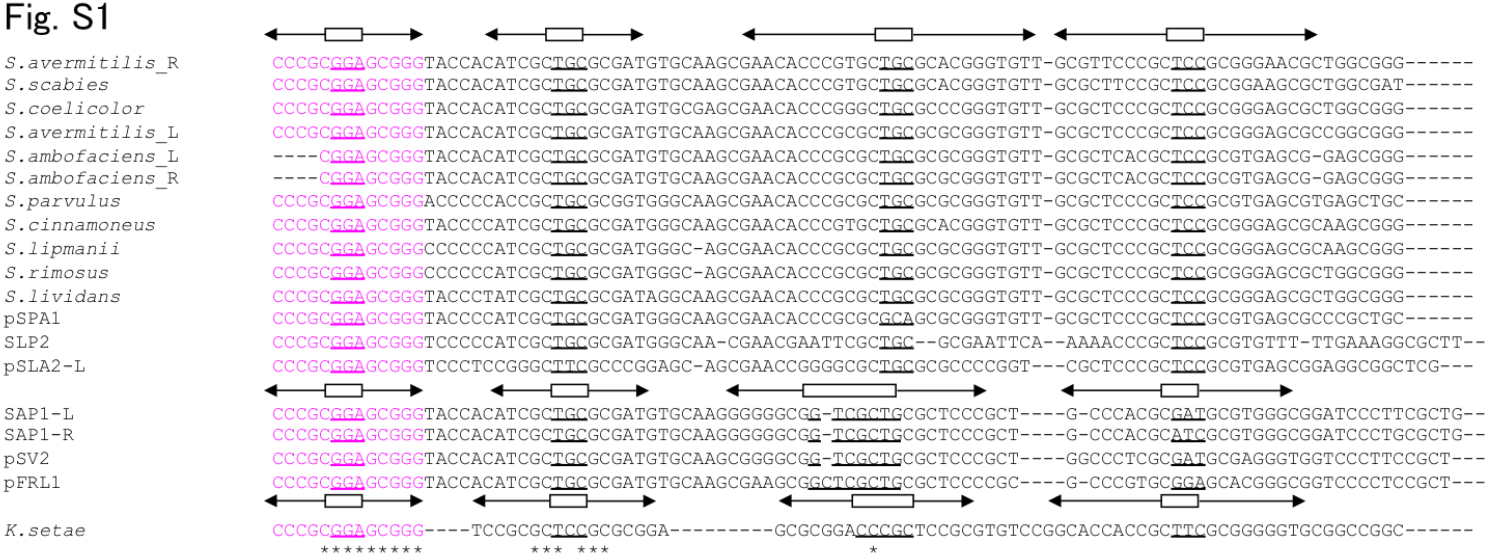


Fig. S1 Comparisons of *K. setae* terminal sequences with *Streptomyces* chromsomes and plasmids in the first 96 bps. Palindrome structures are indicated by arrows and boxes. The first 13 bp conserved among *K. setae* and *Streptomyces* are colored magenta.

The following sequences were retrieved from GenBank : *S. coelicolor* A3(2) (AL939104); *S. avermitilis* (BA000030); *S. cinnamoneus* (AY044253); *S. lividans* (X77096); *S. lipmanii* (AF038454); *S. rimosus* (AY043328); *S. ambofaciens* left arm (AJ937741); right arm (AJ937740); *S. parvulus* pSPA1 (AF038455); *S. lividans* SLP2 (AY225511); *S. rochei* pSLA2-L (AB088224); *S. avermitilis* SAP1 (AP005645); *S. violaceoruber* pSV2 (AY211023); *Streptomyces sp*. FR1 pFRL1 (DQ322651). The genome sequence of *S. scabies* were obtained from StrepDB (http://strepdb.streptomyces.org.uk)

Table S1. List of species used in multi-locus phylogenetic tree analysis

Species
Escherichia coli K12
Kitasatospora setae NBRC 14216
Corynebacterium diphtheriae NCTC 13129
Corynebacterium efficiens YS-314
Corynebacterium glutamicum ATCC 13032
Corynebacterium glutamicum R
Corynebacterium jeikeium K411
Mycobacterium avium 104
Mycobacterium avim subsp. *paratuberculosis* K-10
Mycobacterium gilvum PYR-GCK
Mycobacterium leprae TN
Mycobacterium marinum M
Mycobacterium smegmatis MC2 155
Mycobacterium sp. JLS
Mycobacterium sp. KMS
Mycobacterium sp. MCS
Mycobacterium bovis AF2122/97
Mycobacterium bovis BCG str.
Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis H37Ra
Mycobacterium tuberculosis F11
Mycobacterium ulcerans Agy99
Mycobacterium vanbaalenii PYR-1
Nocardia farcinica IFM 10152
Rhodococcus jostii RHA1
Acidothermus cellulolyticus 11B
Frankia alni ACN14a
Frankia sp. CcI3
Frankia sp. EAN1pec
Kineococcus radiotolerans SRS30216

Tropheryma whipplei TW08/27

Tropheryma whipplei str. Twist

Clavibacter michiganensis subsp. *michiganensis* NCPPB 382

Clavibacter michiganensis subsp. *sepedonicus*

Leifsonia xyli subsp. *Xyli* str. CTCB07

Arthrobacter aurescens TC1

Arthrobacter sp. FB24

Kocuria rhizophila DC2201

Renibacterium salmoninarum ATCC 33209

Salinispora arenicola CNS-205

Salinispora tropica CNB-440

Nocardioides sp. JS614

Propionibacterium acnes KPA171202

Saccharopolyspora erythraea NRRL 23338

Streptomyces avermitilis NBRC 14893

Streptomyces coelicolor A3(2)

Streptomyces griseus subsp. *griseus* IFO 13350

Streptomyces scabiei 87.22

Catenulispora acidiphila DSM 44928

Thermobifida fusca YX

Bifidobacterium adolescentis ATCC 15703

Bifidobacterium longum NCC2705

Rubrobacter xylanophilus DSM 9941

Table S2. Conserved amino-acid sequences used in multi-locus phylogenetic analysis.

annotation	COG_ID	HAMAP_ID
Predicted GTPase, probable translation factor	COG0012	-
Phenylalanyl-tRNA synthetase alpha subunit	COG0016	MF_00281
Ribosomal protein S12	COG0048	MF_00403
Ribosomal protein S7	COG0049	MF_00480
Ribosomal protein S2	COG0052	MF_00291
Ribosomal protein L11	COG0080	MF_00736
Ribosomal protein L1	COG0081	MF_01318
Ribosomal protein L3	COG0087	MF_01325
Ribosomal protein L22	COG0091	MF_01331
Ribosomal protein S3	COG0092	MF_01309
Ribosomal protein L14	COG0093	MF_01367
Ribosomal protein L5	COG0094	MF_01333
Ribosomal protein S8	COG0096	MF_01302
Ribosomal protein L6P/L9E	COG0097	MF_01365
Ribosomal protein S5	COG0098	MF_01307
Ribosomal protein S13	COG0099	MF_01315
Ribosomal protein S11	COG0100	MF_01310
Ribosomal protein L13	COG0102	MF_01366
Ribosomal protein S9	COG0103	MF_00532
Seryl-tRNA synthetase	COG0172	MF_00176
Ribosomal protein S15P/S13E	COG0184	MF_01343
Ribosomal protein S17	COG0186	MF_01345
Ribosomal protein L16/L10E	COG0197	MF_01342
Ribosomal protein L15	COG0200	MF_01341
Preprotein translocase subunit SecY	COG0201	-
DNA-directed RNA polymerase, alpha subunit/40 kD subunit	COG0202	MF_00059
Ribosomal protein L18	COG0256	MF_01337
Leucyl-tRNA synthetase	COG0495	MF_00049
Ribosomal protein S4 and related proteins	COG0522	MF_01306

Valyl-tRNA synthetase	COG0525	MF_02004/MF_02005
Metal-dependent proteases with possible chaperone activity	COG0533	MF_01445

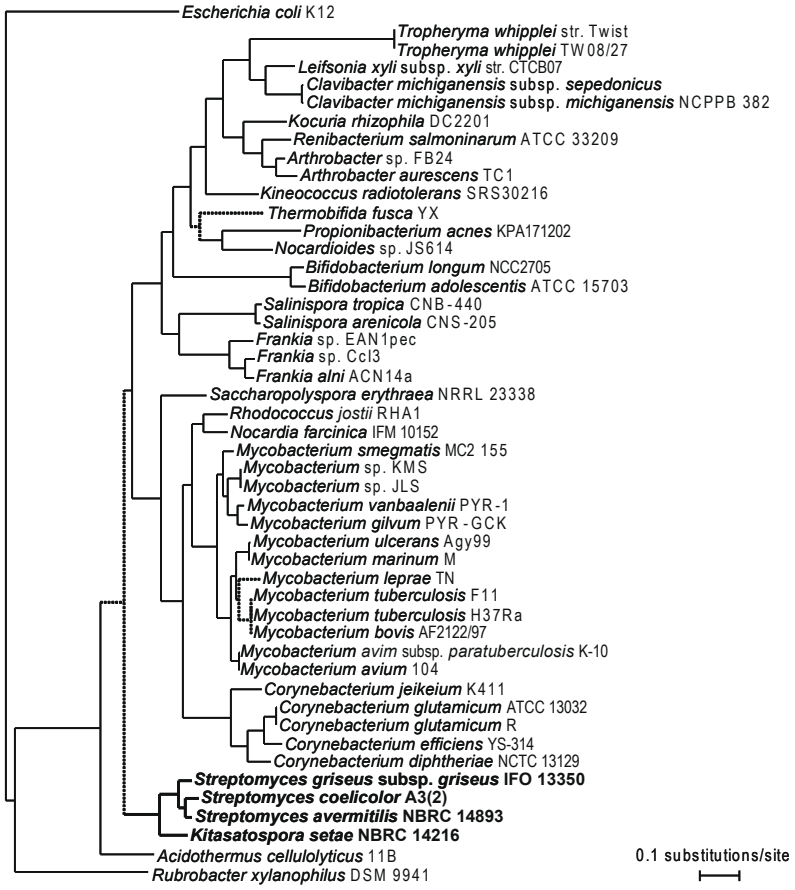


Fig. S2

Phylogenetic tree based upon amino acid sequences of 31 protein-coding genes analyzed by Maximum-likelihood method. Branches with less than 90% bootstrap support are represented in dashed lines. Lists of organisms and genes used for the analysis are shown in supplementary materials Table S1 and S2, respectively. Names of the organisms mentioned in the text are shown in bold type.

CHAPTER-II

Development of secondary metabolite biosynthetic gene cluster database

1. Introduction

In the research of biosynthesis clusters, PKS and NRPS has mainly focused on because of their major roles in constructing complex carbon frameworks. A complex carbon structure is assembled sequentially from simple carbon building blocks, such as acyl-CoA and amino acids. The extension of each carbon unit is catalyzed by a set of functional domains, collectively termed as a ‘module’, encoded in a PKS and a NRPS. A minimal set of domains functional as a ‘module’ in a PKS comprises ketosynthase (KS), acyltransferase (AT) and acyl carrier protein (ACP) domains. The chemical structure of each starter/extender carbon unit can be predicted by examining substrate-specificity determining residues of the AT domain and the presence of optional ketoreductase (KR), dehydratase (DH) and enoylreductase (ER) domains (53–56). Similarly, a minimal set of modules of an NRPS comprises condensation (C), adenylation (A) and peptidyl carrier protein (PCP) domains. The specificity for each starter/extender amino acid is determined by active residues in the A domain, and loaded amino acids are modified by optional domains, such as methyltransferase (MT), epimeration (E) and reductase (R) domains (57). Therefore, precise identification of functional domains and assignment of substrate specificity give important information for the elucidation of the biosynthetic mechanism. Furthermore, as the production of

compounds is thought to be orchestrated work of enzymes encoded within the biosynthesis cluster, immense knowledge of functional roles of tailoring enzymes, transcriptional regulators and transporters also give important clues for the characterization of biosynthesis pathway.

Since the first elucidation of the PKS gene cluster for erythromycin in the early 1990s (58–60), many gene clusters responsible for the biosynthesis of polyketide compounds have been reported in research papers and corresponding international nucleotide sequence databases collection (INSDC) entries (DDBJ/GenBank/EMBL) (61, 62). However, newer analysis data on previously uncharacterized genes or additional sequence data that extend the cluster region are rarely incorporated into existing INSDC entries and are not stored in a comprehensive manner. Each scientist has to read and search a chain of reference citations to find out INSDC entries related to the biosynthetic cluster and to accumulate related knowledge.

Here, we developed a database of biosynthetic clusters of secondary metabolites called DoBISCUIT. Currently, 62 known polyketide biosynthetic clusters are registered. Each biosynthetic cluster is manually curated in terms of sequence collection, reference collection and gene annotation. The database mainly consists of cluster information pages with related references and gene description pages. Our database enables easy access to comprehensive information related to biosynthetic clusters, which also serves as a useful reference to facilitate investigation of secondary metabolite synthesis.

2 Database content

2.1 Data sources and sequence collection

The core data of DoBISCUIT is based on INSDC entries describing each biosynthetic cluster of a known bacterial secondary metabolite. Data collection started from a comprehensive review of literature that reported discoveries of biosynthetic clusters. Articles were collected from PubMed with search term ‘biosynthesis cluster’. The corresponding INSDC accession numbers were extracted from these articles or by searching GenBank by the name of each compound. Further literature collection has been achieved using a paper recommendation system, PubMedScan (http://medals.jp/pubmedscan/), which automatically reports articles highly related to a collection of literature.

In some cases, subsequent investigation revealed the existence of additional functional genes adjacent to a previously identified gene cluster. As a result, the entire biosynthetic cluster was divided into multiple INSDC entries registered separately. For instance, the entire biosynthetic cluster of Jadomycin B comprised five INSDC entries. Positional relationships of previously identified genes and newly found genes are described only in the original article and not assigned in each INSDC entry. In such cases, we collected all INSDC entries comprising the biosynthesis cluster (as far as we could determine) and reconstructed the whole cluster in DoBISCUIT as described in the paper. Currently, we released 62 gene clusters comprising 91 INSDC entries (Table 3).

2.2 Reference collection

Three major phases in the investigation of a biosynthetic cluster may be required; 1) identification of bioactive secondary metabolites, 2) sequencing of the corresponding biosynthetic gene cluster and 3) functional characterization of each gene in the cluster. This process can, therefore, produce multiple research articles. For example, there are at least 37 research articles related to the biosynthetic cluster of Actinorhodin. In DoBISCUIT, we collected as many references as possible by extensive PubMed searches using the compound name, organism name and gene name as search terms. Currently, we have collected 377 references for 62 biosynthesis gene clusters (Table 3).

2.3 Gene descriptions

Gene descriptions in each INSDC entry vary considerably, depending on the submitters. As is often the case, a gene product is represented by an abbreviation, such as capitalized gene symbol. Such a description gives little information about the function of the gene. In many other cases, different submitters use different vocabulary to describe the same gene product. For developing a useful database, it is important to provide intelligible information for users. In DoBISCUIT, genes are described in a controlled vocabulary after manual curation based on experimental data in reference articles and similarities to known proteins. If there are further analysis results published after the release of biosynthetic cluster sequences, the gene descriptions are updated to represent their latest function. Reference information is assigned to each gene to provide evidence for the annotations.

When reviewing many biosynthetic clusters, we also encountered cases where the function of a formerly uncharacterized gene could be inferred from similarities to characterized genes found in other gene clusters. In these cases, we annotated the gene based on the similar genes after careful evaluation of similarities. Of 1695 genes assigned in INSDC entries, we updated 1366 gene descriptions collectively (Table 3).

According to our updated annotations, genes encoded in biosynthetic clusters are expected to play various biological roles, including previously unrecognized roles. We classified all genes based on their biological role. The functional distribution of genes encoded in biosynthetic clusters is shown in Table 4.

As PKS and NRPS play central roles in constructing the carbon backbone, we also emphasized their annotations. Their module and domain assignments were determined by considering the structure of the synthesized compound and the results of domain searches. By careful examination of biosynthetic cluster information, an understanding of synthetic process of secondary metabolites can be gained. Supplementary information concerning what moieties and small molecules the genes would produce is also assigned in DoBISCUIT.

2.4 User availability and web interface

The main content of DoBISCUIT comprises information pages of the biosynthetic cluster (Cluster information page) and each gene encoded in the biosynthetic cluster (CDS information page). Users can also search the contents of DoBISCUIT by keywords, module structures, and sequence similarities.

On the top page, users are presented with the main table of biosynthetic clusters, with multiple options for viewing their attributes. By selecting any of the biosynthetic clusters, users can view a Cluster information page.

2.5 Cluster information page

This page shows integrated information about the biosynthesis cluster (Fig. 4a) and has six sections: compound, origin, genomic map, PKS/NRPS modules, references and data download. The compound section displays the chemical structure, biological activities and various structural attributes, such as chain length and sugar attachment. The origin section displays the bacterial strain name that produces the compound. Users can follow a hyperlink to access the culture collection distributing the strain. The original INSDC entries of the biosynthetic cluster are also displayed. The genomic map section displays the coordinates of genes encoded in the biosynthetic cluster. If the biosynthetic cluster is represented by multiple INSDC entries, they are merged into a single map and the relative location of each entry is displayed on the map. Each gene is colored based on its biological function. The PKS/NRPS modules section displays the domain organization of each module in these enzymes. The deduced substrate of each AT or A domain is shown in the right-most column. Inactive domains are shown in lowercase letters. The reference section displays collected references concerning the biosynthetic cluster, with hyperlinks to PubMed records. The data download section allows users to download certain types of data files: nucleotide sequence of the cluster, CDS nucleotide/amino acid sequences in multi-FASTA format, and curated annotations

in CSV format or GenBank format.

A list of CDSs encoded by the biosynthetic cluster is displayed in another tab of the page (Fig. 4b). CDSs are ordered based on the relative position in the biosynthesis cluster. The list includes a summary of the annotation, including product name, gene name, keyword, and functional category of each gene.

2.6 CDS information page

This page shows integrated information about each CDS in the biosynthetic cluster (Fig. 5). This page has six sections: location, annotation, genomic map, PKS/NRPS modules, sequence and features. The location section displays basic information about the CDS, such as position, length, source organism and INSDC entry. The annotation section displays functional information. Functional category, product name (in controlled vocabulary) and other notes assigned by the annotators are displayed. The original product name and gene name assigned in INSDC entries are also displayed side-by-side. References and corresponding UniProt entries are presented as the evidence of the annotation. If the gene was annotated based on similarities to other sequences, identifiers and hyperlinks for these similar sequences are displayed. The sequence section displays the nucleotide and amino acid sequences of the CDS. The displayed sequence can be switched between nucleic acid and protein by tab buttons. In the case of PKS/NRPS, each domain region is highlighted by a different color. Signature sequences of AT and A domains are also highlighted, with their respective substrates displayed in balloons (Fig. 5a). The feature section displays the results of

automatic searches by bioinformatics tools. “Show BLAST table” button has a hyperlink to the result of a similarity search (BLASTP) (19) executed against the UniProt database (21). Domain assignments obtained by InterProScan (22) are also displayed.

2.7 Search menus

Various search menus are provided in DoBISCUIT. A simple text search form is provided in the upper-right corner of all pages and other search menus can be accessed by following the links in the upper-left panel.

In the simple text search, the search target is restricted to frequently used fields, i.e. compound name, organism name, product name and gene name. Search results are separately presented under ‘Cluster’ and ‘CDS’ tabs.

In the text search menu, users can execute more detailed searches within DoBISCUIT by entering search keywords, specifying the target fields and selecting the target clusters. Target gene clusters can be selected by their attributes, such as PKS type, attached sugar and chain length. Spaces between words are regarded as an 'AND' search term. The search result will be displayed as a list of clusters matching the search conditions. The hyperlink can be followed to a particular biosynthesis cluster page, or compound(s) of interest can be selected. Pressing the CDS tab permits browsing the CDSs.

We also provide a module search menu to find PKSs and NRPSs containing a particular domain composition within the modules. All of the module patterns registered in DoBISCUIT are displayed in the upper part of the menu. Alternatively, auxiliary

input boxes in the middle part of the menu can be used to specify the composition. The result of a module search displays a list of CDSs containing the entered domain composition.

To search homologous CDSs in DoBISCUIT, a BLAST utility is also provided. We provide several kinds of BLAST databases; cluster (containing the whole cluster sequence), CDS (containing all assigned CDSs) and domain (containing all domains assigned in CDSs). The BLAST search results are displayed separately as a list (left) and as an alignment (right). Clicking the 'B' button in the list part displays the alignment calculated by the bl2seq program (19, 63), and clicking the 'T' button displays the alignment calculated by the T-COFFEE program (64).

3 Discussion

3.1 Use of DoBISCUIT in genome mining

The number of genome projects is rapidly growing because of advances in sequencing technologies and decreasing costs. As of the summer of 2012, 103 genome projects intended for the genus *Streptomyces* are registered in the Genomes Online Database (GOLD) (65). Perhaps many of these genome projects are intended to discover or investigate secondary metabolites produced by *Streptomyces* bacteria (66–68). Effective *in silico* identification of biosynthetic clusters from genome sequences is thought to be essential, and very useful web tools have been published (69–71). These web tools identify domains contained in PKS/NRPS and propose similar known biosynthetic clusters to their own. However, in the next stage of genome mining, users

will discover that the information cannot be obtained efficiently from suggested INSDC entries. DoBISCUIT can provide functional annotation of each gene and a comprehensive collection of references. Using a module search, users can obtain a list of CDSs containing the same domain composition as their own. It could be more appropriate to use finely curated small database as a reference than searching a vast amount of patchy information.

3.2 Use of DoBISCUIT in combinatorial biosynthesis

Combinatorial biosynthesis approaches have been attracting attention for the generation of novel natural products and for the production of non-natural derivatives (72, 73). Using recently developed gene manipulation technology, heterologous expression of biosynthetic clusters has been established in *E. coli* (74–76) and *Streptomyces avermitilis* (77). Genetic modification of biosynthetic clusters and/or introduction of a particular mutation also offer opportunities to obtain derivatives of original metabolites. DoBISCUIT can provide functionally classified lists of known biosynthetic cluster genes, which will enable users to easily identify candidates with specific activities for modification of the biosynthetic process. The CDS information page also provides detailed information on what reaction each gene product catalyzes. Users will be able to judge potential applicability of genes to their combinatorial biosynthesis project.

3.3 Use of DoBISCUIT in assessing novelty of biosynthetic genes

The KS domain and the A domain are essential constituents in PKS and NRPS, respectively; therefore, the phylogenetic relationships of their sequences are well related to chemical structures of final products, sequencing analysis of these domains has been often used for assessing the potential of microorganisms to produce novel secondary metabolites (78, 79). In such studies, PCR amplification of KS and A domains, followed by cloning and sequencing, is conventionally used, and the novelty of each domain is assessed by similarity to known domain sequences. DoBISCUIT provides a curated set of domain sequences of known biosynthesis cluster genes for a BLAST database, allowing users to judge the novelty of their sequences by BLAST searches. However, only limited numbers of PKS and NRPS genes corresponding to known bioactive compounds have been identified so far. To fill the gap between large numbers of polyketide/nonribosomal peptide metabolites and only limited numbers of identified biosynthesis gene clusters, we massively sequenced KS and A domains of 464 type strains of the genus *Streptomyces* and 333 antibiotic-producing actinomycetes preserved at NBRC (Komaki *et al.* in preparation). Currently, the resultant nucleotide and amino acid sequences of more than 18,000 domains are also available in DoBISCUIT. BLAST searches against this data set will complement the above approach for assessing the novelty of biosynthetic genes in strains that users collected and try to analyze. In addition, novelty indices of KS/A domain sequences assigned in each NBRC strain (Fig. 6) will help users to select bacterial strains suitable for their research purposes, because it represent not only the novelty but also abundance of PKS/NRPS genes in each strain.

4 Future Perspectives

DoBISCUIT (http://www.bio.nite.go.jp/pks/) enables easy access to comprehensive information related to biosynthesis clusters and to form a standard reference for their investigation. The content of DoBISCUIT will be updated with new biosynthesis clusters and new findings on existing genes. Although the current version of DoBISCUIT mainly focuses on PKS and NRPS, secondary metabolites also comprise other compounds, such as thiopeptides, aminoglycosides and terpenoids. We aim to collect a wider range of biosynthetic clusters in terms of organisms and compounds in future versions of DoBISCUIT.

Our experience of curating a number of biosynthetic clusters allowed us to predict novel functions of previously uncharacterized genes located within the clusters. For example, some uncharacterized genes within the Chalcomycin, Megalomicin and Pikromycin clusters were deduced to encode helper proteins of glycosyltransferases. Further accumulation of cluster information and associated knowledge may help to understand the functions of hitherto unclassified genes.

Table 3. Number of sequences and references registered in DoBISCUIT [a]

data type	Number
gene clusters	62
collected INSDC sequences	91
collected references	377
assigned genes	1695
description changed from original INSDC entry	1366
description accepted	140
description not concerned [b]	189

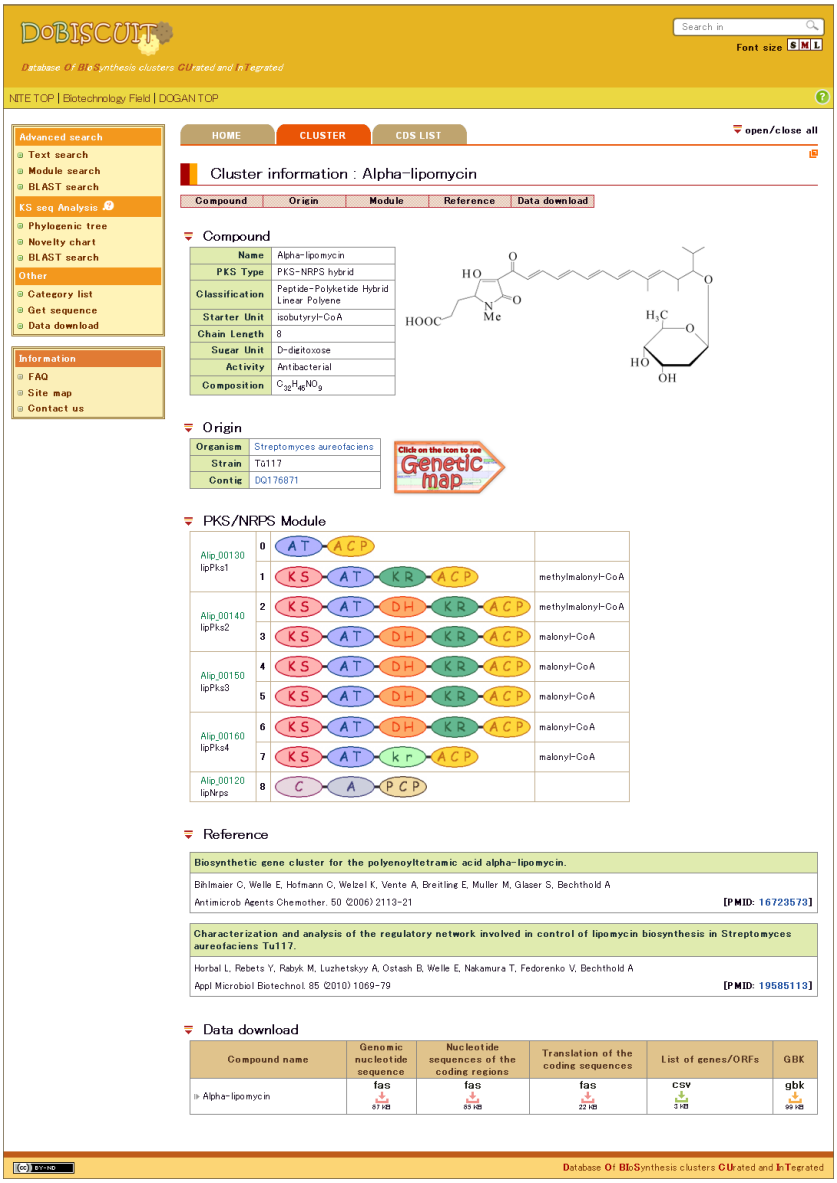
[a] Based on database release as of Aug. 4 2012.

b Genes proved be uninvolved in biosynthesis process.

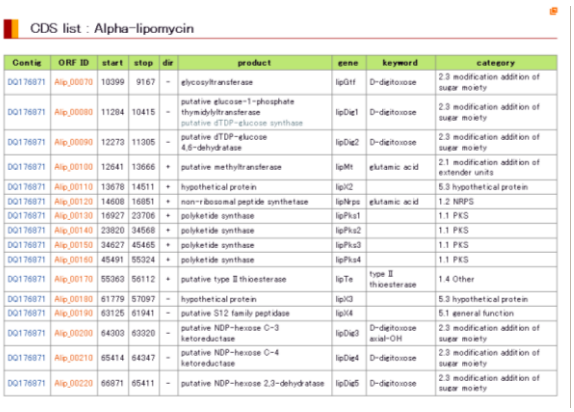
Table 4. Functional categories and number of classified genes

Functional category	Number of CDS
1. Aglycon biosynthesis	
1.1 PKS	275
1.2 NRPS	24
1.3 PKS/NRPS hybrid	4
1.4 Other	49
2. Biosynthesis, modification and addition of aglycon units and moiety	
2.1 extender unit	71
2.2 starter unit	53
2.3 sugar unit	185
3. Modification	
3.1 hydroxylation	2
3.2 methylation	21
3.3 reduction	120
3.4 other modification	88
4. Other function	
4.1 transcriptional regulator	123
4.2 translation	2
4.3 transport	71
4.4 resistance	10
4.5 electron carrier	10
4.6 biosynthesis of butyrolactone	4
5. Putative and unknown function	
5.1 general function prediction	160
5.2 function unknown	37
5.3 hypothetical protein	197

A



B



CDS list : Alpha-lipomycin

Contig	ORF ID	start	stop	dir	product	gene	keyword	category
DQ176871	Alip_00070	10399	9167	-	glycosyltransferase	lipGtf	D-digitoxose	2.3 modification addition of sugar moiety
DQ176871	Alip_00080	11284	10415	-	putative glucose-1-phosphate thymidylyltransferase putative dTDP-glucose synthase	lipDig1	D-digitoxose	2.3 modification addition of sugar moiety
DQ176871	Alip_00090	12273	11305	-	putative dTDP-glucose 4,6-dehydratase	lipDig2	D-digitoxose	2.3 modification addition of sugar moiety
DQ176871	Alip_00100	12641	13666	+	putative methyltransferase	lipMt	glutamic acid	2.1 modification addition of extender units
DQ176871	Alip_00110	13678	14511	+	hypothetical protein	lipX2		5.3 hypothetical protein
DQ176871	Alip_00120	14608	16851	+	non-ribosomal peptide synthetase	lipNrps	glutamic acid	1.2 NRPS
DQ176871	Alip_00130	16927	23706	+	polyketide synthase	lipPks1		1.1 PKS
DQ176871	Alip_00140	23820	34568	+	polyketide synthase	lipPks2		1.1 PKS
DQ176871	Alip_00150	34627	45465	+	polyketide synthase	lipPks3		1.1 PKS
DQ176871	Alip_00160	45491	55324	+	polyketide synthase	lipPks4		1.1 PKS
DQ176871	Alip_00170	55363	56112	+	putative type II thioesterase	lipTe	type II thioesterase	1.4 Other
DQ176871	Alip_00180	61779	57097	-	hypothetical protein	lipX3		5.3 hypothetical protein
DQ176871	Alip_00190	63125	61941	-	putative S12 family peptidase	lipX4		5.1 general function
DQ176871	Alip_00200	64303	63320	-	putative NDP-hexose C-3 ketoreductase	lipDig3	D-digitoxose axial-OH	2.3 modification addition of sugar moiety
DQ176871	Alip_00210	65414	64347	-	putative NDP-hexose C-4 ketoreductase	lipDig4	D-digitoxose	2.3 modification addition of sugar moiety
DQ176871	Alip_00220	66871	65411	-	putative NDP-hexose 2,3-dehydratase	lipDig5	D-digitoxose	2.3 modification addition of sugar moiety

Figure 4.

A. Default view of the cluster information page for the Alpha-lipomycin biosynthetic gene cluster. Information related to chemical compound, producing organism, gene coordinates, domain/modules, related references and downloadable flat files are displayed.

B. CDS list view of the Alpha-lipomycin biosynthetic cluster. The list page is reached via the cluster information page by clicking ‘CDS list’ tab. Relative coordinate of CDSs in biosynthetic cluster, annotated gene descriptions and functional categories are listed.

CDS information: Tatmc_00090

Location | Annotation | Module | Sequence | Feature

Location

Organism	Streptomyces sp.
Strain	CK4412
Compound name	Tautomycetin
Contig	DQ983361
Start / Stop / Direction	38,119 / 60,981 / +
Location	38119..60981
Type	CDS
Length	22,863 bp (7,620 aa)

Click on the icon to see Genetic map

Annotation

Category	1.1 PKS	
Product	polyketide synthase	
Product (GenBank)	tautomycetin biosynthetic PKS	
Gene		
Gene (GenBank)	tmcB	
EC number		
Keyword		
Note		
Note (GenBank)	polyketide type I; tmcB	
Reference	ACC	A4KCE5
	PmId	[17379718] Isolation of the biosynthetic gene cluster for tautomycetin, a linear polyketide T cell-specific immunomodulator from Streptomyces sp. CK4412. (Microbiology. , 2007)
		[20623733] Biochemical and structural characterization of the tautomycetin thioesterase: analysis of a stereoselective polyketide hydrolase. (Angew Chem Int Ed Engl. , 2010)
Related Reference	ACC	C6ZCS0
	PmId	[19191560] Characterization of the tautomycetin biosynthetic gene cluster from Streptomyces griseochromogenes provides new insight into dialkylmaleic anhydride biosynthesis. (J Nat Prod. , 2009)

PKS/NRPS Module

6	KS - AT - DH - ER - KR - ACP	methylmalonyl-CoA
7	KS - AT - DH - ER - KR - ACP	malonyl-CoA
8	KS - AT - DH - KR - ACP	ethylmalonyl-CoA not conserved XAGHS
9	KS - AT - KR - ACP - TE	malonyl-CoA

Sequence

PEP | NUC

```
>polyketide synthase [tautomycetin biosynthetic PKS]
MPSPAGLGRPTTVPTRLFCSTPTSTTSSTSRFAISERPTRELMMAVSVDKLTEALRASLL
ANTRLRQENKTLADAAGEPIAIVGMACRYPGGVAGPDDLWRLVESGGDGVGPFPADRGWD
LDELLGTAESPPASVTSEGGFLADAGDFDAAFFRMSPREALATDPQQRLLLETSWEALEH
AGIDPVALKGSRTGVFAGAYQMGYTGLVTESDEQLRGHAFTGGAGSVTSGRVAYTLGLEG
PAVSVDTACSSSLVAMHLAAQALRAGECTLALAGGVTVMATPETFIAFTVQGGTAADGRC
KSFAEAADGAGWSEGVGMLVLERLSDARRNGHEVFAVMRSSAVNQDGASNGLTAPNGPSQ
QRVIRQALAAAGLSPTDVDAVEAHGTGTKLGDPIEAQALLSTYGHNRPDDRPLWLGSLKS
NIGHTQAAAGVGGVIKMVQALRHGILPRTLHVDRPTTHVEWEQGDVRLLTESIPWPETGR
PRRAGVSSFGVSGTNAHVILEAATADIDAEPAAEGDEPAADGAEVPIALGDFGADGVVPW
VLSGKTDEAVQAQAARLLEHLESVPALRPVDVGWSLAGFRSAFDHRAVVVGAGPDELTAG
LRALAAGRPAAEVLQGSTRGKRDAVFVFPGQGAQWVGMAQELIESSPVFASSMRECAEAL
EPFVDWSLADVLGDEALQRVDVVQPALWAVMVSLAAVWQSAGVSPSAVVGHSQGEIAAA
CVAGALSLSDGARVVALRSRAIAEHLTGGGMMALSVPAAAARDLIADRPGVSLAAINGPS
SVVVSGEADALDSLQEYCEEQGIQARRIAVDYASHSHLVERIEDRLLADLAPVTPRSSSV
PVFSSVTGALMDTAEWDAGYWYRNLRNTVLFEQALTA[illegible]  [6] methylmalonyl-CoA 812..816
PVVGTLRRGEGGLLRLVLALAQAWTHGAPVRWDAVFA[illegible]
ATSPADVAAAGLSPAEHPLL[illegible]
GMLADA[illegible]HRERLG[illegible]LVPMRP[illegible]AALQR[illegible]DDETGLV[illegible]DLDWARFA[illegible]FTAAR
PSPLLSDLPEAAVALVQPDAASTERPALRLAGLSAAEREEALLTLVREAVAAVLGHAPGT
QVDDEQPFKELGFDSLTAVELRNVLQTRTGVGLAAGVVFDHPDPRSLAKALAAALRPADH
AEPQEAQSDYFGELFLQAMRTGELAQAQQLMAGAAQLRLKYGDPAGPEAVPEIVRLGRGQ
LGPQLILVCPTVMTTGPQVYSRLAEELDAGRRVSALVPPGFHGGQALPATLTVLVRSLAD
VVQAEVADGEFALAGHSSGGVVAYEVARELEARGLAPRGVVLIDSYSFDGDGGRPEELFR
SALNERFVEYLRLTGGGNLSQRITAQVWCLELLRGWRPEGLTAPTLYVRPAQPLVEQEKP
EWRGDVLAAMGQVVEAPGDHFTIIEGEHVASTAHIVGDWLREAHAHYSTEGWGGGLRTEE
```

Feature

BLASTP Database:UniProtKB:2010_04	show BLAST table	
HAMAP	No significant hit	
InterPro Database:interpro:26.0	IPR001031 Thioesterase (Domain)	
		PF00975 Thioesterase
	IPR001227 Acyl transferase domain (Domain)	
		G3DSA:3.40.366.10 Ac_transferase_reg
	IPR002364 Quinone oxidoreductase/zeta-crystallin, conserved site (Conserved_site)	
		PS01162 QOR_ZETA_CRYSTAL
	IPR006162 Phosphopantetheine attachment site (PTM)	
		PS00012 PHOSPHOPANTETHEINE
	IPR006163 Phosphopantetheine-binding (Domain)	
		PF00550 PP-binding
		[illegible]M00827 PKS_AT
	IPR020806 Polyketide synthase, phosphopantetheine-binding (Domain)	
		SM00823 PKS_PP
	IPR020807 Polyketide synthase, dehydratase domain (Domain)	
		SM00826 PKS_DH
	IPR020841 Polyketide synthase, beta-ketoacyl synthase region (Region)	
		SM00825 PKS_KS
	IPR020842 Polyketide synthase/Fatty acid synthase, KR (Domain)	
		SM00822 PKS_KR
	IPR020843 Polyketide synthase, enoylreductase (Domain)	
		SM00829 PKS_ER
SignalP		Bacteria, Gram-positive
TMHMM	No significant hit	

Figure 5.

Default view of the CDS information page for a gene encoded in the Tautomycetin biosynthetic cluster. A hyperlink to the CDS information page is provided in the cluster information page and the CDS list page. Manual curation results, such as annotated product name, original product name, gene name and functional category are displayed in the Annotation section. Experimental evidence for this gene is also provided in the reference column of the Annotation section. Functionally important domains for formation of the polyketide chain are displayed in the PKS/NRPS module section. In the sequence, different domains are indicated by different colors.

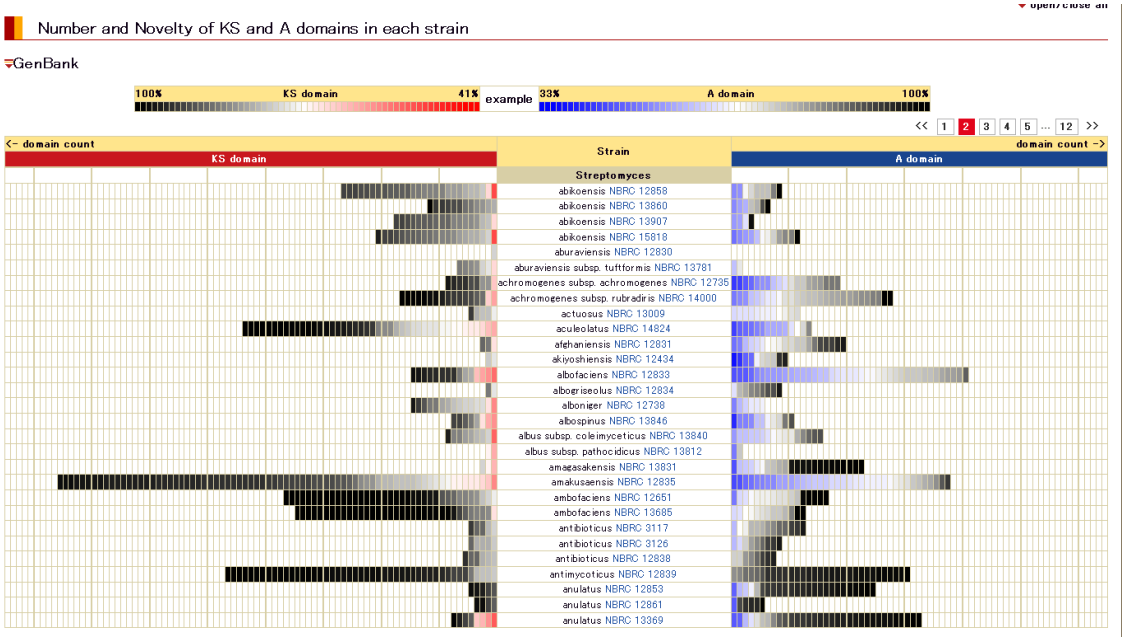


Figure 6.

Novelty chart view of the KS domain and the A domain. Number and similarity to known sequences of KS (left) and A (right) domain sequences amplified from NBRC strains are displayed in a bar chart. The width of the horizontal rectangle represents the number and color gradient of each rectangle represents the degree of similarity for each NBRC strains shown in the center.

CHAPTER-III

Comparative analysis of type-I polyketide synthase and nonribosomal peptide synthetase gene clusters in genus *Kitasatospora*

1 Introduction

The development of rapid and inexpensive sequencing technology resulted in explosion of the number of genome sequences. Since the first release of the complete genome sequence of *K. setae*, the genus *Kitasatospora* came to be recognized as rich reservoir of natural products. At this time, 12 strains consisted of 7 species in the genus *Kitasatospora* are available. Half of the strains are sequenced in massive genome projects aimed to find the novel secondary metabolites from microbes with secondary metabolite production potential. Two strains are sequenced in comparative genome projects of soil-habiting acidophilic actinobacteria. Two other strains are sequenced to understand evolution of biosynthesis clusters isolated from extreme environments such as Himalaya (51). Previously, we have sequenced and analyzed PKS and NRPS gene clusters found in representative seven *Nocardia* strains (80). We have found that the number of PKS and NRPS gene clusters in *Nocardia* strains varies substantially depending on species.

To assess conservation and variety of biosynthesis clusters encoded in *Kitasatospora* genome, we compared PKS and NRPS gene clusters based on domain organization and

amino acid sequence homology. We also predicted functions of genes encoded in biosynthesis clusters using DoBISCUIT database. Despite large number of biosynthesis clusters discovered in *Kitasatospora* genomes, comparative analysis revealed that no biosynthesis cluster is shared more than 3 species. It confirmed that *Kitasatospora* strains are an attractive source of natural compounds and further analyses will broaden the diversity of natural products.

2 Material and Methods

2.1 Acquisition of whole genome sequences

The complete genome sequence of *K. setae* KM-6054 was sequenced in Chapter-I. The genome sequence is also available at DDBJ under the accession number AP010968. The draft genome sequences of *K. setae* NRRL B-16185, *K. azatica* KCTC 9699, *K. cheerisanensis* KCTC 2395, *K. medicidica* KCTC 9733, *K. papulosa* NRRL B-16504, *K. phosalacinea* NRRL B-16228, NRRL B-16230, *Kitasatospora* sp. MBT66, MBT63 and NRRL B-11411 were downloaded from GenBank with accession numbers JNWY01000001-JNWY01000298, JQMO01000001-JQMO01000003, JNBY01000001-JNBY01000178, JQLN01000001-JQLN01000007, JNYQ01000001-JNYQ01000099, JNYE01000001-JNYE01000340, JNWZ01000001-JNWZ01000227, JAIY01000001-JAIY01000045, JAIZ01000001-JAIZ01000847 and JOGH01000001:JOGH01000368, respectively.

2.2 Analysis of PKS-I and NRPS gene clusters

The genome sequence of *K. setae* KM-6054 was analyzed as described in Chapter-I. The draft genome sequences were annotated by prokaryotic genome annotation pipeline and available at RefSeq database (81). As the annotation of *K. medicidica* KCTC 9733, *Kitasatospora* sp. MBT66 and MBT63 were not available, ORFs of these sequences were assigned using Prodigal (82). Annotated ORFs were further searched for signature domains of PKS-I and NRPS genes using the InterPro database (83). We also used antiSMASH (84) for assigning gene clusters and predicting substrates for adenylation domains. We used the cluster assigned by antiSMASH and classified as PKS-I and/or NRPS for further analyses. Function of ORFs encoded in the clusters were deduced by searching DoBISCUIT database described in Chapter-II.

2.3 Search for orthologous gene clusters among species and strains

ORFs having Ketosynthase (KS) domain (IPR014030, IPR014131, IPR020841) or Condensation (C) domain (IPR001242) were identified and these domain sequences were extracted. The amino acid sequences of KS and C domain were aligned using ClustalX and these homology relationships were visualized as dendrogram using NJplot. We considered *Kitasatospora* genes homologous grouped in monophyletic node in molecular phylogenetic analysis, and also when their domain organizations have high similarity. We also performed BLASTP search against the non-redundant protein sequence database (81).

3 Results and Discussion

3.1 Genomic features of *Kitasatospora* genome

The genomic features of *Kitasatospora* genomes are represented in Table 5. The genome size ranged between 5.7 and 10.3 Mb is compatible to those of *Streptomyces* strains (5.0 – 11.9 Mb). The genomic features of two strains of *K. setae* are almost similar in genomic size, contents of GC bases. On the other hand, two strains in *K. phosalacinea* are different ~1Mb in genome size and their predicted biosynthesis clusters are substantially different. The numbers of the total biosynthesis clusters increased proportionally to the genome size. There are differences in the proportion of abundant biosynthesis clusters type, no NRPS biosynthesis cluster was predicted in *K.azatica* while only one PKS type-I gene cluster was predicted in *K. arboriphila* and *K. phosalaconea* NRRL B-16230. The second column from the right on Table 5 lists predicted gene clusters other than PKS type-I and NRPS, such as terpenes, bacteriocins, siderophores and pigments. The numbers of biosynthesis clusters other than PKS-I and NRPS also illustrates that *Kitasatospora* strains are potential producer of various compounds.

Figure 7 shows PKS-I, NRPS and PKS-I/NRPS hybrid gene clusters shared in *Kitasatospora* genomes. Presumptive orthologous clusters are aligned in the same row of the table. Surprisingly, no gene cluster is shared among all analyzed *Kitasatospora* strains. At a maximum, the gene cluster #1 and #13 are found in 5 strains belonging 3 species. As we have reported that 7 gene clusters are common among the seven strains in *Nocardia* (80), suggesting that gene clusters are less conserved in *Kitasatospora*

strains. No orthologous gene cluster is found in *K. arboriphila* with other *Kitasatospora* strains.

3.2 Intra-species variation of clusters

There are two intra-species strains from *K. setae* and *K. phosalacinea*, respectively. In *K. setae* strains, 8 out of 9 gene clusters of KM-6054 listed in Fig. 7 are shared with NRRL B-16185. The homology of each corresponding gene is around the range of 95 % to 100 % in amino acid sequence identity. On the other hand, only 4 out of 6 and 7 gene clusters in NRRLB-16230 and B-16228 are common in two strains of *K. phosalaconea*. The homology of each corresponding gene is around 80 to 90 % in amino acid sequence identity. Concerning the history of the *K. setae* strains, these two strains are derived from the same lineage. Omura *et al.*, isolated KM-6054 and deposited it to two culture collections, IFO and ATCC, these strains were subsequently inherited to NBRC and NRRL, respectively. Contrary, two strains in *K. phosalacinea* have distinct origin; NRRL B-16230 was isolated from Yunnan Province in China (85), while NRRL B-16228 was isolated by Takahashi *et al.* in Japan (86). It is noteworthy that there are strain-specific gene clusters in KM-6054 and NRRL B-16185 in spite of their same origin. It suggests that there are intra-species variations in biosynthesis clusters despite of their lineage.

3.3 Comparison of setamycin (bafilomycin B1) biosynthesis clusters

There are 19 gene clusters commonly found in some *Kitasatospora* strains, although

the numbers of shared species in each gene cluster are remained as many as 3. Setamycin (bafilomycin B1) biosynthesis cluster (#5 in Fig. 7) are present in 3 other strains (*K. setae* NRRL B-16185, *K. cheerisanensis* KCTC 2395 and *Kitasatospora* sp. MBT66) than *K. setae* KM-6054. It is consistent that *K. cheerisanensis* possesses setamycin cluster because it was reported to produce plecomacrolide group compound, a group of macrolide which setamycin belongs (87). Figure 8 represents gene order and domain composition of setamycin clusters. In *K. setae* NRRL B-16185, gene clusters are divided in 3 contigs because of their incomplete sequence. At the first part of gene cluster, PKS gene corresponding to KSE_73440 ends in the middle of module 2. The second part is composed of only PKS genes, it also breaks at the end of module 8. We cannot find the module 9 that would connects from the terminus of OO58_RS31340 gene by BLAST search of whole genome. These sequences may be mis-assembled or wrongly assigned to other contigs. At the beginning of the third part, CDS is not properly assigned in public data, we could find the gene corresponding KSE_73470 by homology search against 6-frame protein translated nucleotide database. In *K. cheerisanensis*, gene order is well conserved except for LuxR transcriptional regulator encoded at the end of setamycin gene cluster. Because the LuxR transcriptional regulator is not proved to be involved in regulation of setamycin, it may not affect biosynthesis of setamycin. In *Kitasatospora sp.* MBT66, orthologous gene cluster is divided in two contgis. Though PKS gene and its domain organization is well conserved, genes responsible for starter biosynthesis and post modification are not conserved. At the upstream of gene cluster, where starter biosynthesis genes are encoded in other

strains, two-component sensor kinases are encoded. At the downstream of gene cluster, malonyl transferase and putative CoA-ligase are encoded in *K. setae*. It was predicted to be involved in incorporation of flavensomycinyl moiety of setamycin (50). In *Kitasatospora sp.* MBT66, these genes are replaced by methyltransferase (11_338) and glycosyltransferase (11_337). For further analysis of their functional role in biosynthesis, we searched known genes using DoBISCUIT database. The gene 11_337 shows 61% sequence similarity with *aveBI* encoded in avermectin biosynthesis cluster. AveBI was characterized as glycosyltransferase of broad specificity, to both sugars and aglycons (88). The gene 11_338 has 61% similarity to *elmMI* placed in elloramycin biosynthesis cluster. ElmMI was reported to be involved in methylation of L-rhamnose moiety in elloramycin (89). Therefore, these two genes 11_337 and 11_338 may be involved in addition of some sugar to setamycin aglycon and methylation of the sugar moiety. Concerning the domain composition of PKS genes, strains other than *K. setae* KM-6054 contained extra DH and KR domains at module 7. It will make a change to aglycon structure from ketone side chain to double bond. Here we present that there are minor changes of genes between species and strains, it can be affect the structure of compounds. Although, genome sequencing for finding of novel biosynthesis clusters mainly focus on the novelty of aglycons, these minor differences can be used as modification candidate for drug development.

4 Conclusion

There are 26 validly published species in the genus *Kitasatspora*, and here I presented the comparison of PKS and NRPS gene clusters covering about one fourth of published species. It revealed that quite variable features of biosynthesis clusters among microbes belonging to the genus *Kitasatospora*. We also demonstrated that the differences exist even among the same species and the same origin. As many of secondary metabolites are produced as bioactive agents to eliminate others for microbes' survival, it may make sense that the secondary metabolites are highly diverged among *Kitasarspora* strains. On the other hand, it also confirmed that *Kitasatospora* species possess a notable number of secondary metabolite biosynthesis clusters. It is expected that *Kitasatospora* strains would be a good resource in the search for new useful drugs.

Table 5 Genome size and numbers of secondary metabolites biosynthesis clusters in *Kitasatospora* strains

Strain name	Genome size (Mb)	GC %	Number of contigs	Number of CDS*	Number of clusters				
					PKS-I	NRPS	PKS-I/NRPS hybrid	Other	Total
K. setae KM-6054	8.78	74.2	1	7,566	2	5	2	28	37
K. setae NRRL B-16185	8.56	74.3	298	7,367	7	8	1	30	46
K. azatica KCTC 9699	8.27	71.6	3	(7,223)	5	0	1	33	39
K. arboriphila NRRL B-24581	5.75	64.5	45	5,202	1	6	0	8	15
K. cheerisanensis KCTC 2395	8.04	73.5	178	7,896	4	4	1	25	34
K. medicidica KCTC 9733	8.68	71.9	7	(7,462)	7	5	2	36	50
K. papulosa NRRL B-16504	7.57	71	99	6,422	2	5	2	23	32
K. phosalacinea NRRL B-16230	7.62	74.1	227	6,680	1	8	2	17	28
K. phosalacinea NRRL B-16228	8.61	74.1	340	7,656	7	8	2	23	40
Kitasatospora sp. MBT66	10.38	72.6	45	(9,065)	4	10	4	33	51
Kitasatospora sp. MBT63	9.85	72.9	847	(9,189)	4	30	0	34	68
Kitasatospora sp. NRRL B-11411	8.64	73.9	368	7,634	0	4	1	27	32

* Numbers of CDS in parenthesis are predicted values by softwere.

#	synthase type	*K. setae* KM-6054 KSE	*K. setae* NRRL B-16185 OO58	*K. azatica* KCTC9699 original	*K. chherisanensis* KCTC 2395 KCH	*K. medicidica* KCTC 9733 original	*K. papulosa* NRRL B-16504 IA22	*K. phosalacinea* NRRL B-16230 OO48	*K. phosalacinea* NRRL B-16228 OO57	*Kitasatospora sp.* MBT66 original	*Kitasatospora sp.* MBT63 original	*Kitasatospora sp.* NRRL B-11411 IH44	putative products
1	pks-I	65090 - 65610	RS23625 - RS23665 RS30285 - RS30405	-	66850 - 67150	-	-	RS18140 - RS18305 RS19290 - RS19425	RS09250 - RS09440	-	-	-	
2	nrps	65950 - 66030	RS14125 - RS14160	-	-	5_1649 - 5_1654	-	-	-	-	-	-	Ala
3	pks-I	27200 - 27480	-	-	-	2_0187 - 2_0210	-	-	-	-	146_0007 - 146_0013	-	
4	hybrid	70410 - 70680	RS04910 - RS05035	-	-	-	-	-	-	-	-	-	factumycin
5	pks-I	73270 - 73570	RS10745 - RS10830 RS27165 - (57_0565) RS31340 - RS31350	-	04000 - 04150	-	-	-	-	11_0340 - 11_0348 17_0001 - 17_0004	-	-	setamycin
6	pks-I	-	-	-	-	-	RS0105620 - RS0105625	RS28875 - RS28950	RS22700 - RS22765	-	-	-	
7	pks-I	-	-	-	-	-	(20_0054 - 20_0077)	-	RS28430 - RS28550	-	-	-	
8	pks-I	-	-	3_3023 - 3_3038	-	7_0140 - 7_0161	-	-	-	-	-	-	
9	nrps	-	-	-	-	-	-	RS21210 - RS21220	RS31720 - RS31730	9_0029 - 9_0067	-	-	
10	nrps	22540 - 22700	RS34800 - RS34825	-	-	-	-	-	-	-	-	-	x-x
11	nrps	33300 - 33360	RS2715 - (3_0173)	-	-	-	-	-	-	-	-	-	Leu-x
12	nrps	57920 - 58340	RS08930 - RS09010 RS35960 - RS35965	-	-	-	-	-	-	-	-	-	x-Asn-Gly-x-x-Thr-Phe
13	nrps	75380 - 75450	RS26105 - RS26110	-	-	-	RS0129120 - RS0129125	RS25250 - RS25335	-	-	-	RS0124120 - RS0124160	x-Gln-x-Gln-x-Phe
14	nrps	-	-	-	06120 - 06420	-	-	-	RS07870 - RS07875 (222_00001)	-	683_0002 - 683_0009	-	Ser-x-Ser-x-x-x-Orn-Dhb
15	hybrid	-	-	-	74020 - 74040	-	(12_0316) - RS0126855	-	-	4_0201 - 4_0240	-	-	polyketide-Orn
16	nrps	-	-	-	-	-	-	-	RS20340 - RS20365	-	819_0028 - 819_0045	-	Thr-Val-Ser
17	nrps	-	-	-	-	-	-	RS33560 - RS33565	RS31720 - RS31730	-	-	-	x
18	nrps	-	-	-	-	-	-	RS03740 - RS03775	-	3_0936 - 3_0963	135_0001 - 135_0002	-	Val-Tyr-Orn-x
19	nrps	-	-	-	-	-	-	-	-	16_0374 - 16_0394	318_0002	-	Gly-Asp-Lys-x

∗ Number in parenthes represents CDS_ID originaly identified in antiSMASH.

Fig. 7 PKS-I, NRPS and PKS-I/NRPS hybrid gene clusters shared in *Kitasatospora* strains.

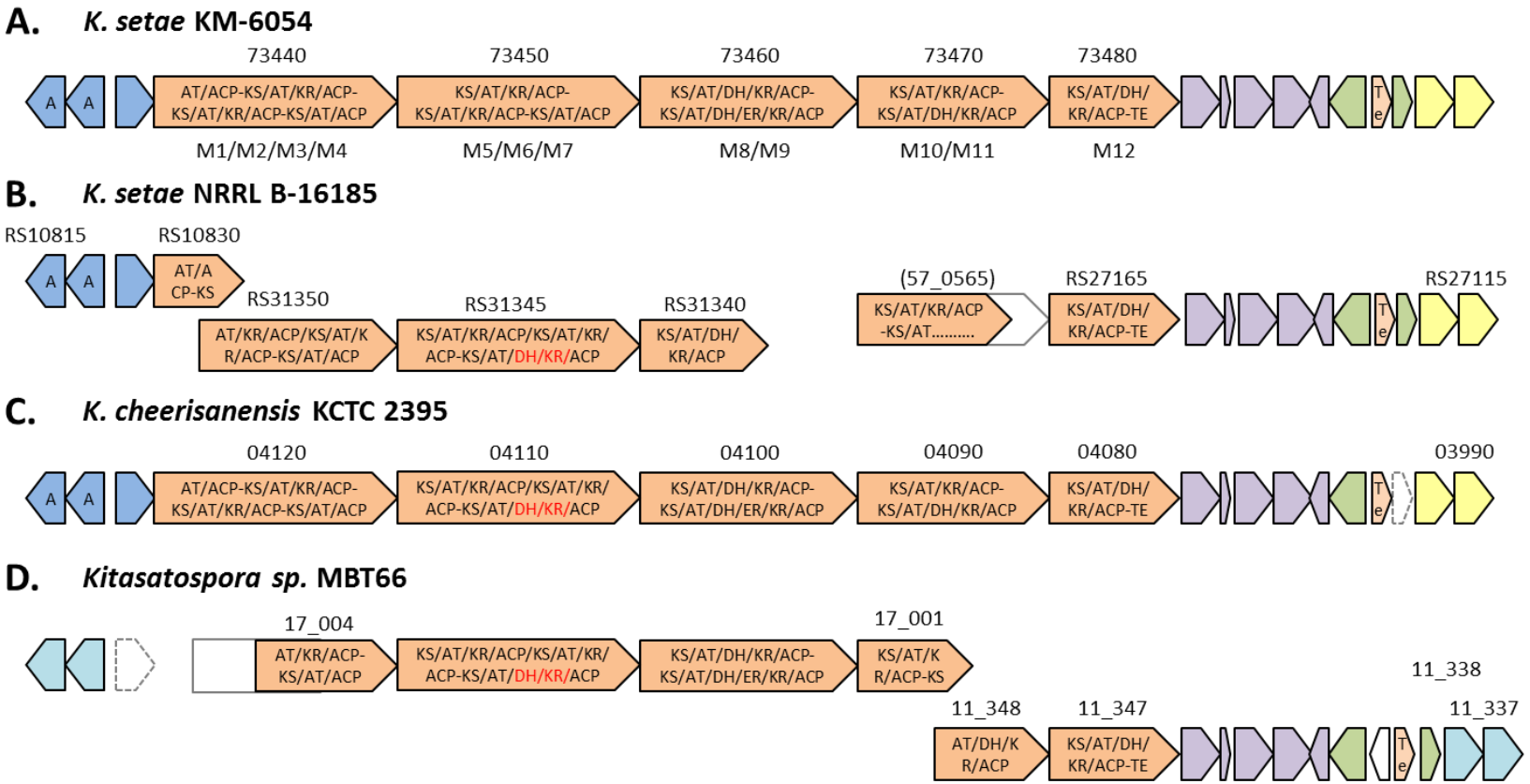


Fig. 8 Comparison of setamycin biosynthesis cluster in *Kitasatospora* strains

Setamycin biosynthesis cluster encoded in *K. seate* KM-5064 (A), *K. setae* NRRL B-16185 (B), *K. cheerisanensis* KCTC 2395 (C) and *Kitasatospora sp.* MBT66 (D). Genes with similar functional roles are marked with the same colors (orange: PKS, purple: synthesis of

attachment moiety, pale blue: synthesis of starter unit, green: transcription, yellow: modification after aglycon formation). Domain structures of each PKS genes are indicated in arrows, and CDS IDs and module numbers are indicated upper and under the arrows.

Conclusion

The long history of genetic and biochemical studies about secondary metabolites brought significant benefits on human life. The genome sequence was one of the epoch-making events in the history of secondary metabolites, uncovering that there are significant number of cryptic biosynthesis clusters in genome sequence.

In Chapter-I, we determined the complete genome sequence of *K. setae* KM-6054. It also revealed that many secondary metabolite biosynthesis clusters, indicating that the importance of the genus *Kitasatospora* as promising source of secondary metabolites.

Many pioneering works about the secondary metabolites were done, however their valuable results were not stored in a comprehensive manner. For more efficient utilization of valuable results, in Chapter-II, we developed novel database that collects various information related to secondary metabolite biosynthesis clusters. The database can be used as useful reference to predict function of genes encoded in biosynthesis cluster.

As an example of comparison analysis of biosynthesis cluster, we compared secondary metabolite biosynthesis clusters found from the genus *Kitasatospora* genomes in Chapter-III. It revealed that the variety of secondary metabolite biosynthesis clusters encoded in these genomes.

Secondary metabolites biosynthesis clusters have been attracting attention for the discovery of novel natural products and for the production of non-natural derivatives. It is expected that our results are of assistance for the further research of secondary metabolites.

References

1. Newman,D.J. and Cragg,G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod*, **75**, 311–335.

2. Park,S.R., Yoo,Y.J., Ban,Y.H. and Yoon,Y.J. (2010) Biosynthesis of rapamycin and its regulation: past achievements and recent progress. *J Antibiot*, **63**, 434–441.

3. Chater,K.F. and Chandra,G. (2006) The evolution of development in *Streptomyces* analysed by genome comparisons. *FEMS Microbiol. Rev.*, **30**, 651–672.

4. Horinouchi,S. (2007) Mining and polishing of the treasure trove in the bacterial genus *Streptomyces*. *Biosci. Biotechnol. Biochem.*, **71**, 283–299.

5. Bentley,S.D., Chater,K.F., Cerdeno-Tarraga,A.M., Challis,G.L., Thomson,N.R., James,K.D., Harris,D.E., Quail,M.A., Kieser,H., Harper,D., *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.

6. 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**, 526–531.

7. Ohnishi,Y., Ishikawa,J., Hara,H., Suzuki,H., Ikenoya,M., Ikeda,H., Yamashita,A., Hattori,M. and Horinouchi,S. (2008) Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.*, **190**, 4050–4060.

8. Momose,I., Sekizawa,R., Hirosawa,S., Ikeda,D., Naganawa,H., Iinuma,H. and Takeuchi,T. (2001) Tyropeptins A and B, new proteasome inhibitors produced by *Kitasatospora sp*. MK993-dF2. II. Structure determination and synthesis. *J Antibiot*, **54**, 1004–1012.

9. Yoon,T.M., Kim,J.W., Kim,J.G., Kim,W.G. and Suh,J.W. (2006) Talosins A and B: new isoflavonol glycosides with potent antifungal activity from *Kitasatospora kifunensis* MJM341. I. Taxonomy, fermentation, isolation, and biological activities. *J. Antibiot.*, **59**, 633–639.

10. Omura,S., Takahashi,Y., Iwai,Y. and Tanaka,H. (1982) *Kitasatosporia*, a new genus of the order *Actinomycetales*. *J. Antibiot.*, **35**, 1013–1019.

11. Wellington,E.M., Stackebrandt,E., Sanders,D., Wolstrup,J. and Jorgensen,N.O. (1992) Taxonomic status of *Kitasatosporia*, and proposed unification with *Streptomyces* on the basis of phenotypic and 16S rRNA analysis and emendation of *Streptomyces* Waksman and Henrici 1943, 339AL. *Int. J. Syst. Bacteriol.*, **42**, 156–160.

12. Ochi,K. and Hiranuma,H. (1994) A taxonomic review of the genera *Kitasatosporia* and *Streptoverticillium* by analysis of ribosomal protein AT-L30. *Int. J. Syst. Bacteriol.*, **44**, 285–292.

13. Zhang,Z., Wang,Y. and Ruan,J. (1997) A proposal to revive the genus *Kitasatospora* (Omura, Takahashi, Iwai, and Tanaka 1982). *Int. J. Syst. Bacteriol.*, **47**, 1048–1054.

14. Sekine,M., Tanikawa,S., Omata,S., Saito,M., Fujisawa,T., Tsukatani,N., Tajima,T., Sekigawa,T., Kosugi,H., Matsuo,Y., *et al.* (2006) Sequence analysis of three plasmids harboured in *Rhodococcus erythropolis* strain PR4. *Environ. Microbiol.*, **8**, 334–346.

15. Takarada,H., Sekine,M., Kosugi,H., Matsuo,Y., Fujisawa,T., Omata,S., Kishi,E., Shimizu,A., Tsukatani,N., Tanikawa,S., *et al.* (2008) Complete genome sequence of the soil actinomycete *Kocuria rhizophila*. *J. Bacteriol.*, **190**, 4139–4146.

16. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

17. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

18. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.

19. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403–410.

20. 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**, 251–253.

21. UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, **40**, D71–5.

22. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L., *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*, **37**, D211–5.

23. Lima,T., Auchincloss,A.H., Coudert,E., Keller,G., Michoud,K., Rivoire,C., Bulliard,V., de Castro,E., Lachaize,C., Baratin,D., *et al.* (2008) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.*, **37**, D471–478.

24. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T., *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–4.

25. Ren,Q., Chen,K. and Paulsen,I.T. (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.*, **35**, D274–9.

26. Rawlings,N.D., Morton,F.R., Kok,C.Y., Kong,J. and Barrett,A.J. (2008) MEROPS: the peptidase database. *Nucleic Acids Res.*, **36**, D320–5.

27. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–4.

28. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

29. Laslett,D. and Canback,B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–16.

30. Worning,P., Jensen,L.J., Hallin,P.F., Staerfeldt,H.H. and Ussery,D.W. (2006) Origin of replication in circular prokaryotic chromosomes. *Environ. Microbiol.*, **8**, 353–361.

31. Ohtsubo,Y., Ikeda-Ohtsubo,W., Nagata,Y. and Tsuda,M. (2008) GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics*, **9**, 376.

32. Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.*, **35**, D343–6.

33. Jakimowicz,D., Majka,J., Messer,W., Speck,C., Fernandez,M., Martin,M.C., Sanchez,J., Schauwecker,F., Keller,U., Schrempf,H., *et al.* (1998) Structural elements of the *Streptomyces oriC* region and their interactions with the DnaA protein. *Microbiology*, **144 (Pt 5)**, 1281–1290.

34. Bao,K. and Cohen,S.N. (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**, 774–785.

35. Huang,C.H., Lin,Y.S., Yang,Y.L., Huang,S.W. and Chen,C.W. (1998) The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. *Mol. Microbiol.*, **28**, 905–916.

36. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science (80-.).*, **311**, 1283–1287.

37. Copeland,A., Lapidus,A., Rio,T.G.D., Nolan,M., Lucas,S., Chen,F., Tice,H., Cheng,J.F., Bruce,D., Goodwin,L., *et al.* (2009) Complete genome sequence of *Catenulispora acidiphila* type strain (ID 139908T). *Stand. Genomic Sci.*, **1**, 119–125.

38. Takahashi,Y. and Omura,S. (2003) Isolation of new actinomycete strains for the screening of new bioactive compounds. *J. Gen. Appl. Microbiol.*, **49**, 141–154.

39. Omura,S., Otoguro,K., Nishikiori,T., Oiwa,R. and Iwai,Y. (1981) Setamycin, a new antibiotic. *J. Antibiot.*, **34**, 1253–1256.

40. Nett,M., Ikeda,H. and Moore,B.S. (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep*, **26**, 1362–1384.

41. Komatsu,M., Tsuda,M., Omura,S., Oikawa,H. and Ikeda,H. (2008) Identification and functional analysis of genes controlling biosynthesis of 2-methylisoborneol. *Proc. Natl. Acad. Sci. USA*, **105**, 7422–7427.

42. Jiang,J., He,X. and Cane,D.E. (2007) Biosynthesis of the earthy odorant geosmin by a bifunctional *Streptomyces coelicolor* enzyme. *Nat. Chem. Biol.*, **3**, 711–715.

43. Cane,D.E., He,X., Kobayashi,S., Omura,S. and Ikeda,H. (2006) Geosmin biosynthesis in *Streptomyces avermitilis*. Molecular cloning, expression, and mechanistic study of the germacradienol/geosmin synthase. *J. Antibiot.*, **59**, 471–479.

44. Weber,T., Laiple,K.J., Pross,E.K., Textor,A., Grond,S., Welzel,K., Pelzer,S., Vente,A. and Wohlleben,W. (2008) Molecular analysis of the kirromycin biosynthetic gene cluster revealed beta-alanine as precursor of the pyridone moiety. *Chem. Biol.*, **15**, 175–188.

45. Tanabe,T., Funahashi,T., Nakao,H., Miyoshi,S., Shinoda,S. and Yamamoto,S. (2003) Identification and characterization of genes required for biosynthesis and transport of the siderophore vibrioferrin in *Vibrio parahaemolyticus*. *J. Bacteriol.*, **185**, 6938–6949.

46. Lynch,D., O'Brien,J., Welch,T., Clarke,P., Cuiv,P.O., Crosa,J.H. and O'Connell,M. (2001) Genetic organization of the region encoding regulation, biosynthesis, and transport of rhizobactin 1021, a siderophore produced by *Sinorhizobium meliloti*. *J. Bacteriol.*, **183**, 2576–2585.

47. Ueki,M., Suzuki,R., Takamatsu,S., Takagi,H., Uramoto,M., Ikeda,H. and Osada,H. (2009) Nocardamin Production by *Streptomyces avermitilis*. *Actinomycetologica*, **23**, 34–39.

48. Choi,S.-U., Lee,C.-K., Hwang,Y.-I., Kinoshita,H. and Nihira,T. (2004) Cloning and functional analysis by gene disruption of a gene encoding a gamma-butyrolactone autoregulator receptor from <i>Kitasatospora setae<i>. *J. Bacteriol.*, **186**, 3423–30.

49. Aroonsri,A., Kitani,S., Hashimoto,J., Kosone,I., Izumikawa,M., Komatsu,M., Fujita,N., Takahashi,Y., Shin-ya,K., Ikeda,H., *et al.* (2012) Pleiotropic control of secondary metabolism and morphological development by KsbC, a butyrolactone autoregulator receptor homologue in *Kitasatospora setae*. *Appl. Environ. Microbiol.*, **78**, 8015–24.

50. Hwang,J.Y., Kim,H.S., Kim,S.H., Oh,H.R. and Nam,D.H. (2013) Organization and characterization of a biosynthetic gene cluster for bafilomycin from *Streptomyces griseus* DSM 2608. *AMB Express*, **3**, 24.

51. Girard,G., Willemse,J., Zhu,H., Claessen,D., Bukarasam,K., Goodfellow,M. and van Wezel,G.P. (2014) Analysis of novel kitasatosporae reveals significant evolutionary changes in conserved developmental genes between *Kitasatospora* and *Streptomyces*. *Antonie Van Leeuwenhoek*, **106**, 365–80.

52. Girard,G., Traag,B.A., Sangal,V., Mascini,N., Hoskisson,P.A., Goodfellow,M. and van Wezel,G.P. (2013) A novel taxonomic marker that discriminates between morphologically complex actinomycetes. *Open Biol.*, **3**, 130073.

53. Jenke-Kodama,H., Borner,T. and Dittmann,E. (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. *PLoS Comput Biol*, **2**, e132.

54. Khosla,C., Kapur,S. and Cane,D.E. (2009) Revisiting the modularity of modular polyketide synthases. *Curr Opin Chem Biol*, **13**, 135–143.

55. Del Vecchio,F., Petkovic,H., Kendrew,S.G., Low,L., Wilkinson,B., Lill,R., Cortes,J., Rudd,B.A., Staunton,J. and Leadlay,P.F. (2003) Active-site residue, domain and module swaps in modular polyketide synthases. *J Ind Microbiol Biotechnol*, **30**, 489–494.

56. Kakavas,S.J., Katz,L. and Stassi,D. (1997) Identification and characterization of the niddamycin polyketide synthase genes from *Streptomyces caelestis*. *J Bacteriol*, **179**, 7515–7522.

57. Stachelhaus,T., Mootz,H.D. and Marahiel,M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol*, **6**, 493–505.

58. Cortes,J., Haydock,S.F., Roberts,G.A., Bevitt,D.J. and Leadlay,P.F. (1990) An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea*. *Nature*, **348**, 176–178.

59. Donadio,S., Staver,M.J., McAlpine,J.B., Swanson,S.J. and Katz,L. (1991) Modular organization of genes required for complex polyketide biosynthesis. *Science (80-.).*, **252**, 675–679.

60. Summers,R.G., Donadio,S., Staver,M.J., Wendt-Pienkowski,E., Hutchinson,C.R. and Katz,L. (1997) Sequencing and mutagenesis of genes from the erythromycin biosynthetic gene cluster of *Saccharopolyspora erythraea* that are involved in L-mycarose and D-desosamine production. *Microbiology*, **143 (Pt 1**, 3251–3262.

61. Kodama,Y., Mashima,J., Kaminuma,E., Gojobori,T., Ogasawara,O., Takagi,T., Okubo,K. and Nakamura,Y. (2012) The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res*, **40**, D38–42.

62. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res*, **40**, D48–53.

63. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*, **174**, 247–250.

64. Di Tommaso,P., Moretti,S., Xenarios,I., Orobitg,M., Montanyola,A., Chang,J.M., Taly,J.F. and Notredame,C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res*, **39**, W13–7.

65. Pagani,I., Liolios,K., Jansson,J., Chen,I.M., Smirnova,T., Nosrat,B., Markowitz,V.M. and Kyrpides,N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, **40**, D571–9.

66. Doroghazi,J.R., Ju,K.S., Brown,D.W., Labeda,D.P., Deng,Z., Metcalf,W.W., Chen,W. and Price,N.P. (2011) Genome sequences of three tunicamycin-producing *Streptomyces* Strains, *S. chartreusis* NRRL 12338, *S. chartreusis* NRRL 3882, and *S. lysosuperificus* ATCC 31396. *J Bacteriol*, **193**, 7021–7022.

67. Qin,S., Zhang,H., Li,F., Zhu,B. and Zheng,H. (2012) Draft genome sequence of marine *Streptomyces sp*. strain W007, which produces angucyclinone antibiotics with a benz[a]anthracene skeleton. *J Bacteriol*, **194**, 1628–1629.

68. Klassen,J.L., Adams,S.M., Bramhacharya,S., Giles,S.S., Goodwin,L.A., Woyke,T. and Currie,C.R. (2011) Draft genome sequence of *Streptomyces sp*. strain Wigar10, isolated from a surface-sterilized garlic bulb. *J Bacteriol*, **193**, 6999–7000.

69. Anand,S., Prasad,M. V, Yadav,G., Kumar,N., Shehara,J., Ansari,M.Z. and Mohanty,D. (2010) SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res*, **38**, W487–96.

70. Medema,M.H., Blin,K., Cimermancic,P., de Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res*, **39**, W339–46.

71. Starcevic,A., Zucko,J., Simunkovic,J., Long,P.F., Cullum,J. and Hranueli,D. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res*, **36**, 6882–6892.

72. Hopwood,D.A., Malpartida,F., Kieser,H.M., Ikeda,H., Duncan,J., Fujii,I., Rudd,B.A., Floss,H.G. and Omura,S. (1985) Production of "hybrid" antibiotics by genetic engineering. *Nature*, **314**, 642–644.

73. McDaniel,R., Thamchaipenet,A., Gustafsson,C., Fu,H., Betlach,M. and Ashley,G. (1999) Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proc Natl Acad Sci U S A*, **96**, 1846–1851.

74. Zhang,H., Wang,Y., Wu,J., Skalina,K. and Pfeifer,B.A. (2010) Complete biosynthesis of erythromycin A and designed analogs using *E. coli* as a heterologous host. *Chem Biol*, **17**, 1232–1240.

75. Horinouchi,S. (2009) Combinatorial biosynthesis of plant medicinal polyketides by microorganisms. *Curr Opin Chem Biol*, **13**, 197–204.

76. Watanabe,K. and Oikawa,H. (2007) Robust platform for de novo production of heterologous polyketides and nonribosomal peptides in *Escherichia coli*. *Org Biomol Chem*, **5**, 593–602.

77. Komatsu,M., Uchiyama,T., Omura,S., Cane,D.E. and Ikeda,H. (2010) Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism. *Proc Natl Acad Sci U S A*, **107**, 2646–2651.

78. Khan,S.T., Komaki,H., Motohashi,K., Kozone,I., Mukai,A., Takagi,M. and Shin-ya,K. (2011) *Streptomyces* associated with a marine sponge *Haliclona sp*.; biosynthetic genes for secondary metabolites and products. *Env. Microbiol*, **13**, 391–403.

79. Komaki,H., Izumikawa,M., Ueda,J.Y., Nakashima,T., Khan,S.T., Takagi,M. and Shin-ya,K. (2009) Discovery of a pimaricin analog JBIR-13, from *Streptomyces bicolor* NBRC 12746 as predicted by sequence analysis of type I polyketide synthase gene. *Appl Microbiol Biotechnol*, **83**, 127–133.

80. Komaki,H., Ichikawa,N., Hosoyama,A., Takahashi-Nakaguchi,A., Matsuzawa,T., Suzuki,K., Fujita,N. and Gonoi,T. (2014) Genome based analysis of type-I polyketide synthase and nonribosomal peptide synthetase gene clusters in seven strains of five representative *Nocardia* species. *BMC Genomics*, **15**, 323.

81. Tatusova,T., Ciufo,S., Fedorov,B., O'Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–9.

82. Hyatt,D., Chen,G.-L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

83. Mitchell,A., Chang,H.-Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S., *et al.* (2014) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, 10.1093/nar/gku1243.

84. Blin,K., Medema,M.H., Kazempour,D., Fischbach,M.A., Breitling,R., Takano,E. and Weber,T. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**, W204–12.

85. Liu,Z., Ruan,J. and Yan,X. (1984) new species of *Nocardiopsis*. *Wei sheng wu hsueh pao = Acta Microbiol. Sin.*

86. TAKAHASHI,Y., IWAI,Y. and OMURA,S. Two new species of the genus *Kitasatosporca*, *Kitasatosporca phosalacinea sp. nov.* and *Kitasatosporia griseola sp. nov*. *J. Gen. Appl. Microbiol.*, **30**, 377–387.

87. Hwang,J.Y., Kim,S.H., Oh,H.R., Cho,Y.-J., Chun,J., Chung,Y.R. and Nam,D.H. (2014) Draft Genome Sequence of *Kitasatospora cheerisanensis* KCTC 2395, Which Produces Plecomacrolide against Phytopathogenic Fungi. *Genome Announc.*, **2**.

88. Zhang,C., Albermann,C., Fu,X. and Thorson,J.S. (2006) The *in vitro* characterization of the iterative avermectin glycosyltransferase AveBI reveals reaction reversibility and sugar nucleotide flexibility. *J. Am. Chem. Soc.*, **128**, 16420–1.

89. Patallo,E.P., Blanco,G., Fischer,C., Brana,A.F., Rohr,J., Mendez,C. and Salas,J.A. (2001) Deoxysugar methylation during biosynthesis of the antitumor polyketide elloramycin by *Streptomyces olivaceus*. Characterization of three methyltransferase genes. *J. Biol. Chem.*, **276**, 18765–74.

Acknowledgement

I would like to thank sincerely Professor Tohru Gonoi, Chiba University, for providing the opportunity to take a doctorate. I would also like to express my sincere gratitude to his patient advices on this thesis.

I would like to thank greatly Dr. Hisayuki Komaki, NITE Bioresource Center, for his support and collaboration of this research.

I would like to thank deeply Dr. Nobuyuki Fujita for providing the opportunity to start and support on this research.

I would like to thank Ms. Sasagawa, Ms. Yamamoto, Ms. Yoshida and co-workers of NITE Bioresource Center for technical support and discussion for this research.

Finally, I must thank my family for their patience and continued support.

DNA research 17巻　第6号
平成２２年１１月８日 公表済
Nucleic Acids Research 41巻
平成２５年1月1日 公表済