Chapter 3

Sequence Analysis

3.1 Introduction:

Sequencing of the type II polyketide synthase cluster from *S. flavisclerotícus* was initiated subsequent to restriction mapping. Firstly, the sequence of *actI* homologous region was determined. For this purpose, the 2.2 kb and 1.7 kb *Bam*HI fragment from the 8.0 kb *Eco*RI fragment were each cloned in to the pBKS vector.

BLAST analysis revealed strong homology to most conserved Ketosynthase (KS α) and Chain Length Factor (KS β), along with a partial sequence of O-methyltransferase gene. The sequencing results from these two fragments not only confirmed that the genomic DNA library constructed was that of PKS II cluster but also encouraged us to dwell further in to the sequence to decipher the complete PKS II cluster. Thus the region adjacent to this 2.2 kb and 1.7 kb *Bam*HI was also cloned and sent for the sequencing.

A combination of two methods - initiating sequencing by universal primers (T3/T7 and/or M13 Forward//M13 Reverse), and continuing sequencing by primer walking was adopted for sequencing most of the DNA. This entails generating short sized and overlapping subclones of the DNA to be sequenced in a vector, pBluescript in this case and getting them custom sequenced.

3.2.1 : Generating subclones for sequencing 8.0 kb *EcoRI* DNA:

The 8.0kb *Eco*RI fragment present in all the four cosmid was first chosen for sequencing as this was the *actI* hybridizing fragment. This fragment was cloned from the cosmid 1.51. The restriction map helped in further subcloning of this 8.0kb *Eco*RI clone into different sized clones. This exercise was mainly done to simplify the sequencing process and most importantly making it economically viable. The 8.0 kb *Eco*RI fragment was sequenced mostly by generating the sub clones like 2.2 kb and 1.7 kb *Bam*HI fragments, 1.6 kb *Kpn*I- *Bgl*II fragment which was the part of 4.0 kb *Bgl*II fragment overlapping the 1.7 kb *Bam*HI fragment and also the 1.7 kb *Bgl*II fragment. These fragments covered ~6.2 kb DNA from the total 8.0 kb *Eco*RI clone. The rest of the DNA from the two ends of the 8 kb DNA was sequenced using T3 and T7 primers.

2.2kb *Bam*HI, 1.7kb *Bam*HI and 1.6 *Kpn*I-*Bgl*II were cloned from the 8.0kb *Eco*RI fragment. The 8.0 kb *Eco*RI fragment was digested with the *Bam*HI enzyme, the digested sample was separated on the 0.8% agarose gel and 2.2kb and 1.7 kb fragments were gel eluted and cloned in a *Bam*HI digested pBKS vector. Similarly 1.9kb *Bgl*II fragment and 1.6 kb *Kpn*I-*Bgl*II were digested with respective enzyme and were cloned in to *Bam*HI digested pBKS and *Kpn*I-*Bam*HI digested pBKS vector respectively (Fig.3.2.1A and Fig.3.2.1B). The cloning of 5.0kb *Eco*RI fragment has already been explained in the previous chapter section 2.1.1.

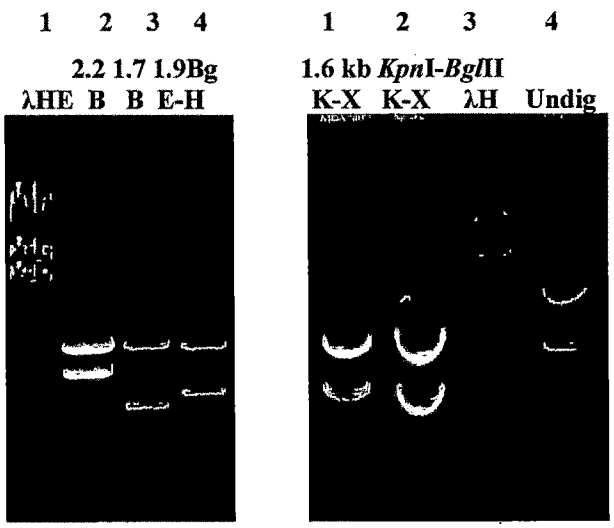


Figure: 3.2.1: Confirmation of the clones by restriction digestion: A . 2.2kb cloning by *Bam*HI (lane2), 1.7 kb cloning by *Bam*HI (lane3) and 1.9 kb *Bgl*II DNA cloning with *Eco*RI and *Hind*III (4). B . *Kpn*I + *Xba*I double digestion of the 1.6 kb *Kpn*I-*Bgl*II clone compared with the undigested DNA.

The gels above represent the cloned fragments validated before they were sent for sequencing. Gel A shows the 2.2 kb *Bam*HI, and 1.7 kb *Bam*HI clones digested with *Bam*HI enzymes (Lane no. 2 and 3) and 1.9 kb *Bgl*II digested with *Eco*RI and *Hind*III enzymes. Gel B shows the cloning of 1.6 kb *Kpn*I-*Bgl*II fragment digested with *Kpn*I and *Xba*I (lane no 1 and 2) and undigested 1.6 kb *Kpn*I-*Bgl*II DNA (lane no 4.).

About 4.2 kb sequence was submitted to NCBI (Nucleotide accession no AY461806).

3.2.2. Analysis of Sequence of the 8.0 kb of PKS DNA:

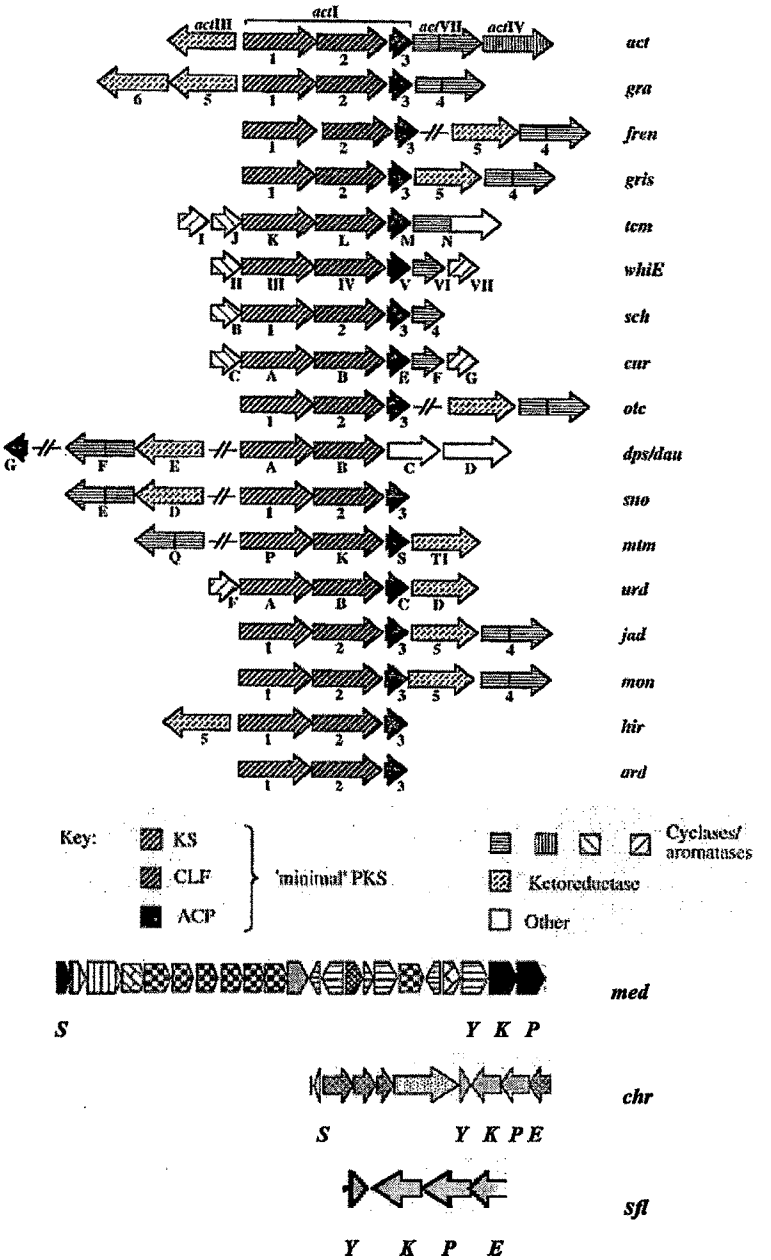
The sequence determined for 8.0 kb *Eco*RI fragement was analysed for the presence of genes responsible for the synthesis of polyketide antibiotic. The organism under study, as has been described earlier was probed with the *actI* genes which codes for the 'minimal PKS' for the actinorhodin cluster. This minimal PKS basically constitutes three essential gens which are responsible for the synthesis of the basic polyketide structure. The genes are namely Ketosynthase (KS), Chain Length Factor (CLF) and Acyl Carrier Protein (ACP). The sequencing results obtained show the presence of KS and CLF but absence of ACP which is very uncharacteristic of the aromatic polyketide synthase gene cluster. This kind of alteration is seen in very few gene clusters like daunomycin, mederrmycin, and enterocin. Along with the KS and CLF genes, *S.flavisclerotícus* polyketide cluster shows the presence of genes which are homologous to genes responsible for glycosylation of the basic backbone molecule. The presence of these genes together suggests that the synthesis of sugar and synthesis of basic structure of polyketide molecule starts simultaneously. These sugars then get attached to the basic backbone structure. 5.0 kb DNA on the other hand contains the genes homologous to NRPS (Non Ribosomal Peptide Synthase) and acyl CoA ligase (see below for details).

That the genes for the NRPS are linked with the PKS cluster and that one end of the cluster has been reached is evident from the analysis.

3.2.2.1: Sequencing result of 8.0 kb *Eco*RI fragment.

8.0 kb *Eco*RI fragment reveals the presence of sequence for the proteins like Ketosynthase, Chain length factor, O-methyl transferase, TDP-4,6 dehydratase and partial sequence for the cyclase. As mentioned earlier, the 8.0 kb fragment; just like duanomycin (Fig.3.2.2.1) and mederrmycin, does not show the presence of the Acyl Carrier Protein (ACP) along with the KS and CLF, instead it marks the presence of the genes which are involved in the terminal modification of the glycan part of the molecule. The organization of the genes in a particular order may give an insight into the kind of synergy between the genes to give an accomplished product. The presence of the genes

responsible for the basic backbone synthesis i.e the minimal PKS genes along with the cyclase indicates that, once the basic structure is synthesized it undergoes the cyclisation process. The presence of these genes together might be indicating to the fact that this kind of organizational arrangement is required to complete a set of operon in order to carry out the synthesis of this kind of complex secondary metabolites. Another striking evidence is the presence of sugar modifying or synthesizing genes along with the genes for the basic backbone synthesis. The genes like O-methyl transferase and TDP- 4,6 dehydaratase in the same direction as the minimal PKS genes is indicative of the fact that the complex sugar synthesis required for the completion of the final structure of the molecule starts simultaneously along with the synthesis of the basic structure and are the part of the same operon.



Key: ***Y*** **: Cyclase,** ***K*** **: Ketosynthase,**

P**: Chain Length Factor,** ***E*****: NDP- 4-6, Glucose dehydratase**

Figure: 3.2.2.1: Gene organization of the Various PKS clusters showing the transcriptional coupling of the KS, CLF and ACP with the exception in *dps/dau*, *med*, *chr* and *sfl*.

Given below is the BLAST analysis for the 8.0 kb sequence. The organism with the maximum homology with the gene has been shown:

O-methyltransferase

Organism	Protein	a.a.	Score	Identity	Accession no.
S. griseus	sugar O-methyltransferase	250	2e-18	88%	CAE17530.1
S. Olivaceus	O-methyltransferase III	261	7e-14	72%	CAD57141.1
S .bikiniensis	putative 3'OH-methyltransferase in D-mycinose	255	2e-13	68%	AAS79452.1
S..roseochromogenes subsp. *oscitans*	O-methyltransferase	277	2e-12	62%	AAN65238.1

Table 3.1: BLAST result for the O-methyltransferase, showing the nearest matches in the database.

TDP-glucose-4,6- dehydratase

organism	Protein	a.a.	Score	Identity	Accession no.
S. griseus	TDP-glucose-4,6-dehydratase	324	6e-81	63%	CAE17525.1
S. cyanogens	NDP-hexose 4,6-dehydratase homolog	326	1e-80	63%	AAD13546
S. rimosus subsp. *paromomycinus*	dTDP-glucose 4,6-dehydratase	317	2e-79	63%	AAF82605
S. avermitilis	dTDP-glucose 4,6-dehydratase	355	2e-78	59%	NP_822121.1

Table 3.2: BLAST result for the TDP-glucose-4,6- dehydratase, showing the nearest matches in the database.

Ketosynthase

organism	Protein	a.a.	Score	Identity	Accession no.
S. griseus	Ketosynthase	422	e-167	86%	CAE17527.1
S. antibioticus	Putative Ketosynthase	422	e-149	77%	AAK06784.1
S. fradiea	Putative ketoacyl synthase	426	e-147	76%	CAA60569.1

Table 3.3: BLAST result for the Ketosynthase, showing the nearest matches in the database.

Chain Length Factor

organism	Protein	a.a.	Score	Identity	Accession no.
S. griseus	CLF	422	e-142	63%	CAE17526.1
S. antibioticus	Putative Ketosynthase	404	e-121	56%	AAK06785.1
S. fradiae	Probable Chain length determinant	408	e-116	54%	CAA60570.1

Table 3.4: BLAST result for the Chain Length Factor, showing the nearest matches in the database.

Cyclase

organism	Protein	a.a.	Score	Identity	Accession no.
S. griseus	cyclase	142	1e-19	76%	CAE17525.1
Streptomyces argillaceus	cyclase	150	1e-10	58%	CAA07761.1
Streptomyces peucetius	daunorubicin biosynthesis enzyme	194	2e-08	58%	AAD04719.1

Table 3.5: BLAST result for the Cyclase, showing the nearest matches in the database.

Putative Protein	Conserved Domain	Description
O - methyl transferase	pfam05711: TylF	This family consists of bacterial macrocin O-methyltransferase (TylF) proteins. TylF is responsible for the methylation of macrocin to producetylosin. Tylosin is a macrolide antibiotic used in veterinary medicine to treat infections caused by Gram-positive bacteria and as an animal growth promoter in the swine industry. It is produced by several *Streptomyces sp.* As with other macrolides, the antibiotic activity of tylosin is due to the inhibition of protein biosynthesis by a mechanism that involves the binding of tylosin to the ribosome, preventing the formation of the mRNA-aminoacyl-tRNAribosome complex.
TDP-glucose- 4-6 Dehydratase	PRK10084: PRK10084	dTDP-glucose 4,6 dehydratase, has a putative role in NAD(P)-binding
Ketosynthase	cd00834: KAS_I_II	KASs are responsible for the elongation steps in fatty acid biosynthesis. KASIII catalyses the initial condensation and KAS I and II catalyze further elongation steps by Claisen condensation of malonyl-acyl carrier protein (ACP) with acyl-ACP.
Ketosynthase	Cd00832: CLF	Chain-length factor (CLF) is a factor required for polyketide chain initiation of aromatic antibiotic-producing polyketide synthases (PKSs) of filamentous bacteria. CLFs have been shown to have decarboxylase activity towards malonyl-acyl carrier protein (ACP). CLFs are similar to other elongation ketosynthase domains, but their active site cysteine is replaced by a conserved glutamine.
Cyclase	Pfam04199: Cyclase	Proteins in this family are thought to be cyclase enzymes. They are found in proteins involved in antibiotic synthesis. However they are also found in organisms that do not make antibiotics pointing to a wider role for these proteins. The proteins contain a conserved motif HXGTHXDXPXH that is likely to form part of the active site.

Table 3.6: Homology to Conserved Domain based on which putative proteins were predicted.

3.2.3: Sequencing result of 5.0 kb *Eco*RI fragment.

As described in the chapter 1 and 2, the 5.0 kb *Eco*RI forms the extreme left region of the ~45 kb region. This fragment is unique to cosmid 1.23, and is separated from the 8.0 kb *Eco*RI fragment by the 11.0 kb and 4.0 kb *Eco*RI fragments (Fig 2.1). Sequencing and analysis of the sequence was carried out precisely to find out if the DNA contains the typeII PKS genes. The information for genes which are directly involved in the aromatic polyketide biosynthesis is clearly absent in this DNA. Instead, this region shows the presence of the genes involved in the Non Ribosomal Peptide Synthesis.

The presence of gene like DUF 894 which codes for permease suggests that the genes for antibiotic resistance or ABC transporters are in the vicinity of that region. DNA sequence matching acyl transferase domain and tRNA synthetase class II domains is also indicative of the region presumably important for NRPS. This domain is the core catalytic domain of tRNA synthetases of the subgroup containing glycyl, histidyl, prolyl, seryl and threonyl tRNA synthetases.

The functionality of the DNA in production of NRPS molecule is being tested, however preliminary results show that the cluster when mutated abolishes production of bioactivity against both *E. coli* and *M. luteus* when used as test organisms. The production of the yellow chromophore, the product of PKS genes (see Chapter 4) is not affected by the mutation, implying that the NRPS cluster is not involved in the formation of conjugated NRPS-PKS complex, a class that is responsible for synthesis of hybrid peptide-polyketide class of molecules. This work is in progress.

The organism with the maximum homology with the gene has been shown below:

Non Ribosomal Peptide synthetase (NRPS)

Organism	Protein	a.a.	Score	Identity	Accession no.
S. cellulosum	non-ribosomal peptide synthetase	125	4e-20	69%	embAM746676.1
S. fradiae	lipopeptide antibiotic A54145 biosynthetic gene cluster	131	2e-17	79%	gbDQ118863.1
Saccharopolyspora erythraea	putative non-ribosomal peptide synthetase	102	3e-15	76%	embAM420293.1
Rhodococcus sp	probable polyketide synthase	80	1e-13	72%	CP000431

Table 3.7: BLAST result for the NRPS, showing the nearest matches in the database

Condesation Domain

organism	Protein	a.a.	Score	Identity	Accession no.
Cyanothece sp. PCC 7424	amino acid adenylation domain protein	400	6e-94	42%	gbEDU13484.1
Nostoc punctiforme	amino acid adenylation domain protein	398	4e-88	42%	gbACC81021.1
Myxococcus xanthus	non-ribosomal peptide synthetase	380	4e-76	42%	gbABF87402.1

Table 3.8: BLAST result for the Condesation Domain, showing the nearest matches in the database

Permease

organism	Protein	a.a.	Score	Identity	Accession no.
Brevibacterium linens BL2	Permeases of the major facilitator superfamily	352	2e-64	42%	ZP_00380080.1
Streptomyces collinus	putative major facilitator family transporter	353	3e-53	38%	emb|CAN89665.1
Streptomyces avermitilis	putative export protein	358	6e-52	35%	dbj|BAB69351.1

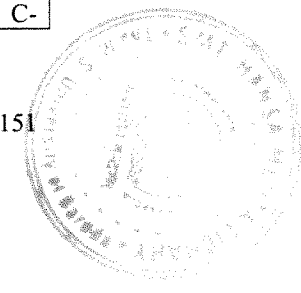
Table 3.8: BLAST result for permease, showing the nearest matches in the database

tRNA synthetase class II core domain

organism	Protein	a.a.	Score	Identity	Accession no.
Ralstonia eutropha	Seryl-tRNA synthetase	158	2e-10	40%	emb|CAJ94821.1|
Bacillus cereus	tRNA synthetase class II	144	2e-08	31%	gb|ABS20546.1|
Bradyrhizobium sp. ORS278]	putative seryl-tRNA synthetase	131	2e-08	37%	emb|CAL80498.1|

Table 3.9: BLAST result for the tRNA synthetase class II core domain, showing the nearest matches in the database

Putative Protein	Conserved domain	description
condensation domain	cl09290	This domain is found in many multi-domain enzymes which synthesise peptide antibiotics. This domain catalyses a condensation reaction to form peptide bonds in non- ribosomal peptide biosynthesis. It is usually found to the carboxy side of a phosphopantetheine binding domain (pfam00550). It has been shown that mutations in the HHXXXDG motif abolish activity suggesting this is part of the active site.
DUF894 domain Permease	pfam05977	Bacterial protein of unknown function (DUF894). This family consists of several bacterial proteins, many of which are annotated as putative transmembrane transport proteins
tRNA synthetase class II core domain	cd00670	Gly_His_Pro_Ser_Thr_tRNA synthetase class II core domain. This domain is the core catalytic domain of tRNA synthetases of the subgroup containing glycyl, histidyl, prolyl, seryl and threonyl tRNA synthetases. It is primarily responsible for ATP-dependent formation of the enzyme bound aminoacyl-adenylate. These enzymes belong to class II aminoacyl-tRNA synthetases (aaRS) based upon their structure and the presence of three characteristic sequence motifs in the core domain. This domain is also found at the C-



		terminus of eukaryotic GCN2 protein kinase and at the N-terminus of the ATP phosphoribosyltransferase accessory subunit, HisZ and the accessory subunit of mitochondrial polymerase gamma (Pol gamma b) . Most class II tRNA synthetases are dimers, with this subgroup consisting of mostly homodimers. These enzymes attach a specific amino acid to the 3' OH group of ribose of the appropriate tRNA
Acyl Transferase domain	pfam00698	Polyketide synthase modules and related proteins [Secondary metabolites biosynthesis, transport, and catabolism

Table 3.10: Homology to Conserved Domain based on which putative proteins were predicted.

3.3 Conclusion:

This chapter discusses the analysis of the sequences for the 8.0 kb and 5.0 kb *Eco*RI fragment. The *actI* hybridizing region in the 8.0 kb *Eco*RI fragment was cloned into pBKS and was custom sequenced by Microsynth, Switzerland. The sequencing results revealed the presence of Ketosynthase (KSα), Chain length Factor (KSβ), NDP-4-6, glucose dehydratase, O-methyltransferase and cyclase in the 8.0 kb *Eco*RI fragment and presence of Nonribosomal Peptide Sythetase genes in the 5.0 kb *Eco*RI fragment. The gene organization seen here was uncharacteristic of the other aromatic PKS containing organism. For example, we find that the gene for the protein involved in the glycan formation like NDP-4-6, glucose dehydratase was transcriptionally coupled with the genes KS and CLF which are responsible for the synthesis basic aglycon molecule. The absence of ACP along with the KS and CLF is a rare observation in the Type II PKS cluster. Daunomycin, medermycin and enterocin are the other three examples with such an exception.