Chapter 7
Databases of polyketide synthases
7.1 INTRODUCTION

In the previous chapters we have reported the computational analysis of various domains and linkers belonging to different classes of polyketide synthases. Our analysis indicated that, in order to correlate the sequence of PKSs to their polyketide products, it is essential that the information on domains and modules of various PKS proteins and their substrate specificities are organized in the form of a searchable database. At present the sequences of various multi-functional NRPS and PKS proteins are available in different sequence databases like SWISSPROT (Bairoch et al., 2004) or GENBANK (Benson et al., 2005). However, the organization of domains and modules in these proteins, and their substrate specificities, has not been comprehensively annotated. The standard domain identification tools like Conserved Domain Database (CDD) search (Marchler-Bauer et al., 2005) are often found to be inadequate for accurate depiction of domain organization in these multi-functional proteins since they fail to detect some of the key functional domains in PKS proteins. This inspired us to develop searchable databases on domain organization and substrate specificity of various experimentally characterized PKS proteins. Based on a comprehensive analysis of various experimentally characterized PKS proteins catalogued in these databases we have developed computational tools for prediction of domain organization and substrate specificity of uncharacterized PKS proteins. These prediction protocols have also been integrated with the databases as appropriate query interfaces. Together they provide a set of user-friendly tools for correlating polyketide chemical structures with the organization of domains and modules in the corresponding modular polyketide synthases. They also allow users to extensively analyze and assess the sequence homology of various polyketide synthase domains, thus providing guidelines for carrying out domain and module swapping experiments.

7.2 METHODS

7.2.1 Organization of databases & search facility

Figure 7.1 depicts the flowchart of the organization of various polyketide synthase databases. The primary source of information for this database is the protein
sequences of PKS, CHS and hybrid NRPS / PKS gene clusters and chemical structures of their experimentally characterized biosynthetic products. Since all type I PKS clusters consist of multi-functional proteins, various functional domains present in each ORF are identified using suitable bioinformatics tools. The accuracy of domain assignment is checked by comparing the computationally assigned domain organization with the experimentally validated domain arrangements, which can be deduced from the chemical structure of the product using the appropriate logic of biosynthesis. Each database also provides appropriate query interfaces for comparing the sequences of various PKSs catalogued in them. These sequence comparison options can provide user-friendly guidelines to design genetic manipulation experiments for biosynthesis of novel polyketides. The databases and query interfaces have appropriate computational tools for extraction of the active site residues and prediction of the starter and extender substrate specificity of each potential acyltransferase domain. In case of CHS and type I iterative PKS, comparison with chemical structure provides information on the number of iterative chain elongation steps carried out by the corresponding protein. The putative active site residues of these domains are identified by alignment of their sequences on the three dimensional structures of homologous proteins. Using the above mentioned data mining protocol, the information on domain organization, sequences of domains and linkers, specificities of domains involved in selection of starters and extenders, their active site residue patterns as well as the chemical structure of the natural product have been stored in three databases. These databases are PKSDB, ITERDB and CHSDB, named as per the type of natural product and mechanism of their biosynthesis. By analyzing the domain organization of a large number of PKS proteins and their active site residue patterns, knowledge based predictive rules have been developed for in silico identification of various PKS domains and assignment of their substrate specificities. The knowledge base obtained from the analysis of these characterized clusters has been used to develop appropriate query interfaces for predicting domain organization and substrate specificity of uncharacterized PKS proteins.
Chapter Seven: Databases of polyketide synthases

7.3 RESULTS

7.3.1 PKSDB

PKSDB is a database of type I modular PKSs and it has been compiled based on a systematic analysis of the sequence and structural features of experimentally characterized modular PKS clusters. It contains information on nineteen modular PKS gene clusters. These are amphotericin (Caffrey et al., 2001), ascomycin (Wu et al., 2000), avermectin (Ikeda et al., 1999), erythromycin (Aparicio et al., 1994), epothilone (Tang et al., 2000), megalomycin (Volchegursky et al., 2000), myxalamide (Silakowski et al., 1999), myxothiazol (Silakowski et al., 2001), niddamycin (Kakavas et al., 1997), nystatin (Brautaset et al., 2000), oleandomycin (Shah et al., 2000), pikromycin (Xue et al., 2000), pimaricin (Aparicio et al., 1999), pyoulteorin (Nowak-Thompson et al., 1997), rifamycin (August et al., 1998), rapamycin (Schwecke et al., 1995), soraphen...
PKSDB gives an intuitive pictorial depiction of domain organization along with the chemical structure of the polyketide product for easy identification of chemical moieties added by various modules. It also provides sequences of domains and linkers, their comparative analysis with homologous domains, active site residues and substrate specificities of AT domains and also information on catalytic activity of reductive domains. SEARCHPKS, a major component of PKSDB, is a powerful software for detection and analysis of PKS domains in a polypeptide sequence. This software can identify various PKS domains from a given query sequence by using an automated computational protocol, and display the results as a pictorial representation of the domain organization. The program also provides appropriate interfaces to carry out a number of different analyses for each of the depicted domains. Figure 7.2 depicts various features of SEARCHPKS.

![Flowchart depicting various features of SEARCHPKS.](image-url)
7.3.1.1 Domain Identification and Depiction of Domain Organization

A major component of SEARCHPKS is the automated computational protocol for correct identification of various PKS domains in a polypeptide sequence. Domain identification is carried out by pairwise sequence alignment of the query sequence with template sequences of KS, AT, DH, ER, KR, TE and ACP domains. Sequences are aligned using a local version of BLAST program (Altschul et al., 1990) downloaded from NCBI (Wheeler et al., 2002). BLOSUM62 scoring matrix and default values for gap penalties are used for sequence alignments and only alignments having E-value less than 0.000001 are considered as statistically significant hits. Template sequences of KS, AT, DH, ER and KR domains have been taken from module 4 of erythromycin, while template for TE is from module 6 of erythromycin. The boundaries of these template sequences are chosen based on the sequence analysis by Donadio & Katz (Donadio and Katz, 1992). It may be noted that various PKS domains identified as per the boundaries suggested by Donadio & Katz have been used in a number of domain swapping experiments. Thus it was considered appropriate to choose template domains based on their work. However, detailed sequence analysis (reported in chapter two) of various experimentally characterized PKS clusters indicated that, using the ACP domain of erythromycin module 4 as template and E-value cut off of 0.000001, several functional ACP domains could not be detected. Therefore, for the identification of the ACP domains, a representative set of 73 diverse sequences of ACP family taken from Pfam database (Bateman et al., 2004) is used as templates in SEARCHPKS program. It may be noted that, the standard method for identification of various functional domains in a protein, is to use CDD search. As discussed in chapter two, our analysis of domain organization in 19 characterized modular PKS clusters has indicated that, CDD search fails to detect any DH domains and cannot distinguish between KR and ER domains. SEARCHPKS correctly detects all the reductive domains by using appropriate DH, ER and KR templates. Correct identification of reductive domains is essential for predicting the chemical structure of the final polyketide product. Therefore, SEARCHPKS is more useful than CDD search for the purpose of identification of polyketide reductive domains and correlating them with their polyketide product.

After the identification of the boundaries of various domains in the query sequence, SEARCHPKS depicts the arrangement of domains and linkers in a pictorial
format with clickable links leading either to their amino acid sequences in FASTA format or for further analysis involving that domain. The modular organization is highlighted by using different colors for different modules in a potential PKS cluster. Since, PKS clusters often consist of multiple ORFs, SEARCHPKS provides options for submitting up to 10 different polypeptide sequences in a single query, so that the domain organization of an entire cluster can be viewed together in a single output. Figure 7.3 shows a typical result from SEARCHPKS for the ORFs Rv1661 and Rv1664 from the genome of \textit{M. tuberculosis} strain H37Rv. As can be seen, the ORF Rv1661 contains a complete module consisting of all the reductive domains, while ORF Rv1664 is a minimal module containing only the KS, AT and ACP domains.

Figure 7.3: A typical use of SEARCHPKS for predicting domain organization in two ORFs Rv1661 and Rv1664 from \textit{M. tuberculosis} H37Rv. On entering a 5-letter name for the query and selecting TWO as the number of ORFs, the program displays a form for submitting the sequences of the two ORFs in FASTA format. Upon submitting the query, the program gives a pictorial depiction of the domain organization in these two ORFs. The domains are depicted as filled circles with names of the domains inscribed in them, while the linker regions are represented as filled or shaded lines.
7.3.1.2 Extraction of sequences and homology assessment

SEARCHPKS provides a very convenient interface for extracting sequences of various domains and linkers in FASTA format. Each of the PKS domains identified in the query sequence can also be compared with other similar domains found in 19 characterized PKS clusters, analyzed in detail in chapter two. These 19 PKS clusters have a total of 182 KS, 188 AT, 108 DH, 29 ER, 165 KR, 192 ACP and 17 TE domains. The sequences of all these domains have been stored in PKSDB, which is accessible from the query interface of SEARCHPKS. Upon selecting the option to obtain sequences of homologous domains, the query domain is aligned with every other characterized domain in PKSDB. All the pairwise alignments are then stored by SEARCHPKS in a temporary directory and the user can retrieve a specified number of sequences which are most similar or most diverse from the query domain. For the ORF Rv1661 from *M. tuberculosis*, Figure 7.4 shows a typical example of the usage of SEARCHPKS for extraction of the sequence of an AT domain in FASTA format and retrieval of 10 AT domains most similar to this AT domain. The homologous domains given in this output are in fact clickable links leading to the alignment of the query domain with each of these domains. The program also provides an option for aligning the query domain with any specific domain from the 19 characterized PKS clusters stored in PKSDB. Figure 7.5 shows an example, where the query AT domain has been aligned with an AT domain in the loading module of epothilone PKS. From the alignment page, the user can go to the page depicting the complete domain organization of the PKS cluster from which the subject domain has been selected as well as the chemical structure of the corresponding polyketide product. For example, on clicking the link to the subject cluster in Figure 7.5, the user can access the page containing the domain organization of epothilone cluster along with the chemical structure of epothilone (Figure 7.6). This page accessed from PKSDB, has also been generated using SEARCHPKS and has been appropriately annotated to compare the predicted domain organization with the experimentally validated domain organization of epothilone cluster (Molnar et al., 2000). This option of SEARCHPKS is useful for finding out the type of polyketide products made by homologous domains in various experimentally characterized PKS clusters.
Chapter Seven: Databases of polyketide synthases

Figure 7.4
Screen dumps showing usage of SEARCHPKS for extracting the sequence of an AT domain in FASTA format and retrieving 10 AT domains most similar to the AT domain in query sequence. On clicking the AT domain of the second ORF in the screen shown in Figure 7.2, the program displays a page with a button leading to sequence of this domain in FASTA format and a form for extracting from PKSDB, a specified number of AT domains most similar or most diverse from the query domain. For these homologous AT domains, the program lists the degree of similarity to the query, their active site motif as well as substrate specificity along with links to their alignment with the query.

Figure 7.5
Screen dumps showing alignment of the query AT domain with the AT domain in the loading module of epothilone PKS. On selecting the pair alignment option in the screen shown in Figure 7.3, the program displays a list of PKS clusters and upon selection of a PKS cluster, the program prompts the user to select the module from which the specified domain is to be chosen for alignment. After getting the required user input, the program displays the alignment page and this page also provides links to the modular PKS cluster from which the subject domain has been selected.
The AT sequence is your AT. Which AT would you like to align this sequence with?

### Figure 7.4

**View Pair Alignment**

The query seq is your AT. Which AT seq would you like to align this seq with?

### Figure 7.5

**Alignment between your AT sequence and "epothin.mod1 AT"**

You have chosen the Epothione gene cluster. It has AT domains in the following modules.
Figure 7.6

Pictorial depiction of domain organization for epothilone biosynthetic cluster. All domains in the same module have been depicted using a single color, while different modules have different colors. In the chemical structure, each chemical moiety has been depicted using the same color as the color of the corresponding module of the epothilone cluster which adds it to the growing polyketide chain during biosynthesis of epothilone. Domains which have been correctly predicted by SEARCHPKS have been represented as solid filled circles, while domains which are not predicted by SEARCHPKS have been depicted as dotted filled circles. Catalytically active domains have been depicted using upper case alphabets, while names of the domains have been inscribed in lower case for catalytically inactive domains. The methyl transferase domains have been manually annotated as the software does not include templates for detection of this domain.
Click the box to view chemical structure of epothilone.

* Click any image below for further details:

epoth001.seq

** APPROX PRODUCT CHEMISTRY: **

(epoth002.seq)

** APPROX PRODUCT CHEMISTRY: **

(epoth003.seq)

** APPROX PRODUCT CHEMISTRY: **

(epoth004.seq)

** APPROX PRODUCT CHEMISTRY: **

Figure 7.6
7.3.1.3 Substrate specificity of AT domains

Since AT domains are known to control the specificity for various starter and extender units during polyketide biosynthesis, SEARCHPKS also extracts the key active site residues of AT domains and based on the pattern of these active site residues it attempts to predict their substrate specificity. For each query AT domain, 13 active site residues are extracted from its alignment with the crystal structure of acyltransferase from *E. coli* FAS (IMLA) (Serre et al., 1994). The choice of these 13 active site residues is based on our detailed sequence analysis and molecular modeling calculations, which indicated that the substrate specificity of AT domain is controlled by only few residues in the active site cavity (reported in chapter two). If the 13 active site residues of a query AT domain show an identical match to the corresponding residues in any AT domain of known specificity in our data set from 19 characterized modular PKS clusters, the query domain is assigned the same specificity as that of the matched AT domain. SEARCHPKS also lists the 13 active site residues as well as the known specificities of AT domains homologous to the query. The 13 active site residues of the AT-domain in Rv1664 show exact match with several malonate specific AT domains (Figure 7.4) and as can be seen from Figure 7.3, this AT domain is predicted to be specific for malonate. Similarly, the AT domain in Rv1661 is predicted to be specific for methylmalonate (Figure 7.3). In case no exact match of 13 active site residues is found for the query domain, SEARCHPKS only lists the 13 active site residues and the user can draw inferences about the substrate specificity by comparing this motif with the active site residues of the homologous domains.

7.3.1.4 Information on the chemical moiety added by a PKS module

For modular PKSs, based on the type of reductive domains and specificity of the AT domain, SEARCHPKS attempts to predict the approximate chemical formulae of the moiety likely to be incorporated by a given module in the query sequence. Since SEARCHPKS uses an automated computational approach for prediction of chemical structure, for computational convenience, the chemical formulae are represented in a seven character symbolic form as shown below (Table 7.1).
Table 7.1

<table>
<thead>
<tr>
<th>Module</th>
<th>Symbolic Representation</th>
<th>Chemical structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>KS-AT-ACP</td>
<td>-(CO -CHR)-</td>
<td><img src="image" alt="Chemical structure" /></td>
</tr>
<tr>
<td>KS-AT-KR-ACP</td>
<td>-(CHO-CHR)-</td>
<td><img src="image" alt="Chemical structure" /></td>
</tr>
<tr>
<td>KS-AT-DH-KR-ACP</td>
<td>-(CH=CR)-</td>
<td><img src="image" alt="Chemical structure" /></td>
</tr>
<tr>
<td>KS-AT-DH-ER-KR-ACP</td>
<td>-(CH2-CHR)-</td>
<td><img src="image" alt="Chemical structure" /></td>
</tr>
</tbody>
</table>

R can be H or CH₃ or any other chemical group depending on whether the AT domain is specific for malonate, methylmalonate or any other substrates. With appropriate tools, these symbolic representations can be converted to conventional chemical structures.

7.3.2 ITERDB

ITERDB is a database of type I iterative polyketide synthases. It contains information on 21 type I iterative PKSs. However, they correspond to 14 different polyketide products as ITERDB has sequences from different organisms corresponding to the same polyketide product. Domain identification has been carried out using the program SEARCHPKS. It is interesting to note that, because of the significant homology between modular and iterative PKS domains, SEARCHPKS has been successful in detecting most of the domains in type I iterative PKSs, even if it uses single template sequences from modular PKS for all KS, AT, ER and KR domains. A single template from an iterative PKS has been used for DH identification. A representative set of 73 diverse sequences of ACP family from Pfam database (Bateman et al., 2004) were used as templates for identification of ACP domains. An additional template had to be used for cyclase domain since it is found in iterative PKSs only. As in PKSDB, the domains identified by the computational protocol were compared with experimentally validated domain organization and discrepancies were appropriately annotated using different symbols for catalytically inactive domains and domains not identified by SEARCHPKS. For each type I iterative PKS, apart from the pictorial depiction of domain organization, ITERDB also provides links to the chemical structure of the polyketide product and to PUBMED for the abstract of the publication describing experimental characterization of the cluster. Figure 7.7 shows a typical
example using the compactin PKS (Abe et al., 2002) from ITERDB. In the chemical structure, the moieties added during different iterative cycles have been depicted using different colors. Based on the chemical structures of the polyketide products, substrate specificities have been assigned to various AT domains and putative active site residues have been identified for each AT domain from their alignment with the crystal structure of acyltransferase from E. coli FAS (1MLA) (Serre et al., 1995). ITERDB is closely integrated with PKSDB and it provides appropriate interfaces for extracting sequences of various domains, their substrate specificity and comparing them with other similar domains in ITERDB as well as PKSDB. Figure 7.8 shows available menus in ITERDB for various types of comparative analysis with a selected domain and a typical example for retrieval of a specified number of modular or iterative AT domains most similar to the AT domain in compactin PKS. As can be seen, the matches from modular PKS are colored differently to distinguish them from matches with domains from iterative PKSs. ITERDB also has a search interface for depiction of domains in a query sequence and their comparison with other iterative PKS domains.

**Figure 7.7: Screen shots of ITERDB** showing the compactin PKS cluster. It has six domains which iterate eight times to result in a nonaketide. The structure button reveals the mechanism of compactin formation. Chemical moieties added during different iterations have been colored differently. The PUBMED button links to the paper describing experimental characterization of the cluster. Filled circles denote accurate identifications by our program while dotted circles represent domains that were not predicted. The patterned circle, as in case of the MT domain, indicates domains for which templates have not been used during the identification procedure. Images of domains and linkers are clickable links leading to further information and analysis.
Figure 7.8
Screen shot of the page obtained by clicking the AT domain in Figure 7.7. The page provides interfaces for extracting its sequence in FASTA format and its pair alignment with any other AT domain in PKSDB or ITERDB. It also shows an example of the usage of Sequence Relatives option to obtain active site residues of 10 closest AT domains from ITERDB or PKSDB. Iterative AT matches are marked with a pink ball and modular ones with a blue ball.

Figure 7.9
This screen shot from CHSDB depicts the chemical structure of the starter, extender and final product for acridone synthase. It also has interfaces for extracting FASTA sequence and 32 active site residues based on 1CGZ. Example of a typical use of CHSDB for comparing the 32 active site residues of acridone synthase with those of other CHS proteins.
Figure 7.8

Figure 7.9
7.3.3 CHSDB

CHSDB is a database of type III polyketide synthases. Since type III PKSs are essentially single domain mono-functional proteins, CHSDB primarily contains information on substrate specificity of these proteins, unlike PKSDB and ITERDB which are databases of domain organization and substrate specificity of PKSs. At present CHSDB has 11 CHS like enzymes from various plant species and 3 bacterial type III PKSs. For each protein, CHSDB provides the sequence, the chemical structures of the starter and extender substrates, information on number of condensation cycles carried out by the enzyme and the chemical structure of the enzyme bound linear polyketide as well as the final cyclized product. The active sites of the CHS family of proteins have been characterized in detail based on the crystal structures of several plant CHS proteins with and without substrates (Jez et al., 2000). A set of 32 amino acids which form the active site of CHS proteins have been categorized into six different groups based on their role in CoA binding, starter group binding, catalysis, cyclization etc. The variations in these 32 residues control the specificity of CHS like enzymes for various starter and extender units as well as different modes of cyclization and thus lead to biosynthesis of diverse end products. These 32 active site residues have been extracted for each of the enzymes in CHSDB from their pairwise alignment with 1CGZ, the crystal structure of alfalfa CHS (Ferrer et al., 1999). CHSDB provides appropriate interfaces for comparing either the entire sequence or these putative active site residues in any selected set of CHS proteins and correlating them to their substrate specificities. Figure 7.9 shows information on acridone synthase (Lukacin et al., 1999) as an example to highlight various features of CHSDB. CHSDB also has a user friendly query interface for analyzing putative type III PKS proteins of unknown specificity. Using this option the query sequence can be aligned with any of the characterized type III PKS proteins in CHSDB using a local version of BLAST and its active site residues extracted from alignment with 1CGZ, can be analyzed to get clues about its substrate specificity. These features of CHSDB make it a valuable tool for analyzing large number of type III PKS sequences found in plant and bacterial genomes. CHSDB can also help in rational design of novel type III PKS proteins with altered specificities.
7.3.4 Integrated query interface for hybrid NRPS-PKS proteins

In hybrid NRPS-PKS gene clusters, polyketide synthases are often found on the same ORF as NRPS proteins or as adjacent ORFs. We have integrated PKSDB ITERDB and CHSDB with NRPSDB, a database of non-ribosomal peptide synthetases (Ansari et al., 2004) to develop NRPS-PKS, a comprehensive resource for bioinformatics analyses of NRPS-PKS megasynthases. These databases work as the backend of NRPS-PKS and provide the knowledge base for predicting domain organization and substrate specificity of uncharacterized NRPS / PKS clusters. Benchmarking on a large set of biosynthetic gene clusters has demonstrated that, apart from correct identification of NRPS and PKS domains, NRPS-PKS can also predict specificities of adenylation and acyltransferase domains with reasonably high accuracy. These features of NRPS-PKS make it a valuable resource for identification of natural products biosynthesized by NRPS / PKS gene clusters found in newly sequenced genomes. The training and test sets of gene clusters included in NRPS-PKS correlate information on 307 open reading frames, 2223 functional protein domains, 68 starter / extender precursors and their specific recognition motifs, and also the chemical structure of 101 natural products from four different families.

7.4 DISCUSSION

We have organized the sequence information on various experimentally characterized PKS gene clusters in the form of several searchable computerized databases which facilitate easy extraction of PKS domains, identification of their catalytic activity, active site residues, substrate specificity etc. from a polypeptide sequence. A unique feature of the integrated databases is that they facilitate correlation of sequence information of biosynthetic proteins to the chemical structure of their products. The combined search facility permits comparison of NRPS / PKS domains in terms of their sequence similarity, substrate specificity and active site motifs. These features make these databases a valuable resource for designing experiments to engineer ‘unnatural’ natural products by providing guidelines for domain / module swapping as well as site directed mutagenesis experiments. This database can also provide leads to decipher the natural products biosynthesized by NRPS / PKS clusters.
found in newly sequenced microbial genomes. All databases described here are available at http://www.nii.res.in/nrps-pks.html.