Chapter One

Review of Literature
1.1 NONRIBOSOMAL PEPTIDE SYNTHETASES

Nonribosomal peptide molecules belong to a diverse group of natural products produced by many microorganisms. These peptides are typically 5 to 15 residue long and exhibit a remarkable array of biological activities. Many of them are pharmaceutically valuable compounds like antibiotics, immuno-suppressants, veterinary agents and agrochemicals e.g. gramicidin-S (antibiotic), enterobactin (siderophore), vancomycin (cell wall biosynthesis inhibitor), cyclosporin-A (immunosuppressant), epothilone (antitumor) etc. This diversity in biological activities is attributed to the remarkable structural diversity observed in these peptides, which is in part due to the incorporation of many unusual nonproteinogenic residues. Many such residues, including amino acids in D-configuration and N-methylated amino acids or a variety of hydroxy acids, have been identified. Further modifications by acylation, glycosylation or heterocyclic ring formation, and variations within the peptide backbone resulting in linear, cyclic or branched cyclic molecules also contribute to the enormous variety of structures within this class of molecules. Figure 1.1 shows some of these nonribosomal products classified based on their biological functions and highlight the essential features that increase their structural diversity to enable them to perform such activities.

The nonribosomal peptides are biosynthesized by large multifunctional proteins called nonribosomal peptide synthetases (NRPSs) through sequential incorporation of amino acid building blocks into the growing polypeptide chain. Typically a number of enzymatic domains present in a single polypeptide chain catalyze various reaction steps during the biosynthesis of nonribosomal peptides. Often a given nonribosomal peptide is biosynthesized by a NRPS cluster consisting of multiple NRPS proteins organized in an operon. These multifunctional enzymes can reach remarkable sizes to synthesize a natural product. For example, Tolypocladium niveum uses a single, 1.6 MDa NRPS containing 41 enzymatic sites to synthesize undecapeptide cyclosporine-A (Weber et al., 1994).

NRPS resembles fatty acid synthase (FAS) and polyketide synthase (PKS) in the modular organization of catalytic domains (Keating and Walsh, 1999). While FAS is essentially involved in primary metabolism in fatty acids biosynthesis for energy storage, PKS synthesize another class of biologically important secondary natural products called polyketides. In fact, NRPS, PKS and FAS, all use carrier protein
homolog for transport of intermediates which shows striking similarity in structure and function. Apart from the large number of NRPS and PKS gene clusters, mixed NRPS/PKS clusters have also been identified in several organisms. Mixed NRPS/PKS gene clusters exemplify how nature combines the modular biosynthesis strategies of NRPS and PKS machinery to further increase the structural/functional diversity of the resulting natural products.

1.2 ORGANIZATION OF NRPSs

NRPSs are organized into coordinated groups of active sites termed modules where each module is responsible for catalyzing one single cycle of polypeptide chain
1. Substrate adenylation by A-Domain
2. Thioester formation — loading of PCP
3. Peptide bond formation by upstream C domain
4. Peptide bond formation by downstream C domain

Figure 1.2: Schematic diagram depicting the enzymatic reaction catalyzed by a minimal NRPS module. Each module is colored differently. Curved line emerging from thiolation (T or PCP for peptidyl carrier protein) domains represents a 20 Å phosphopantetheine (4'PP) cofactor which acts as swinging arm to channel the intermediates substrate to active site of different functional domains. Adenylation (A) domain catalyses the selection and activation of substrate and condensation (C) domain catalyzes peptide bond formation between two PCP-bound activated substrates present on adjacent modules.

The essential set of catalytic domains are adenylation (A) domain responsible for substrate selection and activation through adenosine-tri-phosphate (ATP) hydrolysis (Stachelhaus and Marahiel, 1995), a downstream peptidyl carrier protein (PCP) or thiolation (T) domain for the covalent fixation of the activated substrate as thioester (Stachelhaus et al., 1996), and a condensation (C) domain upstream of the A domain (Stachelhaus et al., 1998) as depicted in Figure 1.2. The C domain catalyzes...
the formation of a peptide bond between two PCP-bound activated substrates present on adjacent modules resulting in elongation of peptide by one amino acid and transfer of the growing chain to the PCP domain in the downstream module. The PCP domain downstream to a C domain possesses an aminoacyl moiety as thioester bond, while the upstream PCP domain contains peptidyl group whose residue number depends on the number of modules present upstream to the C domain. Optional modification domains for substrate epimerization, N-methylation and heterocyclization can extend the basic set of domains within a module. The various modules responsible for biosynthesis of a given NRPS product can remain on a single polypeptide chain or on multiple polypeptides. In most bacterial NRPS systems a thioesterase (TE) domain, present at the C-terminus end of the last module is responsible for the release of the final product into solution. Other catalytic domains like specialized C and reductase (R) have also been implicated in the product release (Keating et al., 2001). The principle of NRPSs is explained using biosynthetic machinery of the branched cyclic decapeptide bacitracin-A in Figure 1.3.

1.3 TYPES OF NRPSs

A collinear relation between the primary sequence of the nonribosomal peptide product and the arrangement of modules in the biosynthetic cluster has been found in a large number of NRPS systems. Such knowledge of assembly-line biosynthesis have been utilized to exchange or swap domains and modules of different specificity to produce nonribosomal peptide products with desired sequence (Doekel and Marahiel, 2000; Mootz et al., 2000; Stachelhaus et al., 1995). However, many NRPS clusters characterized recently show substantial deviations from the linear assembly-line enzymology. Based on the biosynthetic mechanisms, NRPSs have been classified into three major groups (Figure 1.4); linear (type A), iterative (type B) and non-linear (type C) NRPS (Mootz et al., 2002). In linear NRPS, the three core domains are arranged in the order C-A-PCP in an elongation module that adds one amino acid to the growing peptide intermediate. First amino acid is incorporated by initiation module which lacks a C domain and the product is released by the last domain. For linear NRPSs, the sequence of the resulting peptide product is entirely determined by the number and order of the modules [Figure 1.4 (i)]. The iterative NRPSs also have the elongation module as C-A-PCP similar to the linear NRPS, but use their modules or domains more than once in the assembly of a single product. The
Figure 1.3: The model of NRPS machinery illustrated by the biosynthesis of bacitracin. $\textit{bacA}$, $\textit{bacB}$, and $\textit{bacC}$ genes are part of an operon. BacA and Bac3 proteins contain five modules each, while Bac2 contains two modules. Recognition and activation of amino acid substrates via their adenylates are accomplished on the adenylation domains before the activated amino acids become concatenated to the cofactor 4'-phosphopantetheine which is part of each module's PCP domain. The stepwise N- to C-terminal assembly of the peptide is shown. The reactions are performed by the condensation domains or, alternatively, by the cyclization domains forming thiazoline heterocycles. Intermediate epimerization domains convert appropriate amino acids into D-configuration (positions 4, 7, 9, and 11). The peptide is cleaved from the enzyme by the action of a C-terminal thioesterase-like domain. Various domains are depicted as circles. Name of each domain is written on the circles, except for A domain for which the specificity is mentioned. The final structure of the product bacitracin-A is shown. Figure adapted from Schwarzer et al., 2003.
Figure 1.3
resulting product thus consists of repeated regions as seen in enterobactin synthetase in [Figure 1.4 (ii)]. In such iterative NRPS systems, \( TE \) domain plays a major role in oligomerization and the final release of the peptide product. The nonlinear NRPS in contrast to linear and iterative NRPS employ complex biosynthetic mechanism as can be seen from [Figure 1.4 (iii)]. Examples of some of the products of these three types of NRPS are listed in Table 1.1.

### 1.4 FUNCTIONAL DOMAINS OF NRPS

The domains which are essential for incorporation of amino acids by an elongation module are \( A \) domain for substrate recognition, a \( PCP \) domain for transfer of the activated substrate to different catalytic centers, and a \( C \) domain for catalyzing the peptide bond formation. \( TE \) domain is required for final release of the product. In addition to these essential domains, there are optional domains which are responsible for various chemical modifications of the nonribosomal peptide backbone. Some of these optional domains are integral part of NRPS while others are separate enzymes. Various features of these domains are discussed below.

#### 1.4.1 The Core Domains

##### 1.4.1.1 The Adenylation Domain

Adenylation domain is typically about 550 amino acids and an integral domain of a NRPS module. As mentioned earlier, \( A \) domains play an important role in specific selection of the cognate substrate from the large pool of available substrates.
Figure 1.4: Various biosynthetic strategies of NRPS.

(i) Linear NRPS; ACV synthetase, the final product ACV, contains three residues which correspond to three modules and the residues are collinear to the specificity of the module. (ii) Iterative NRPS; enterobactin synthetase, three Dhb-Ser-4'PP intermediates are generated on the two modules of the enterobactin NRPS and are oligomerized and cyclized on the TE domain. (iii) Nonlinear NRPS; vibriobactin synthetase, the siderophore is synthesized from three molecules of Dhb (a), two molecules of Thr (b) and one molecule of soluble norspermidine (c). Figure adapted from Mootz et al., 2002.
Figure 1.4

Linear (type A)

Module 1

Module 2

Module 3

\(a+b+c\)

\(\delta\)-Aminoadipyl-cysteinyld-valine (ACV, the \(\beta\)-lactam precursor)

Iterative (type B)

Module 1

Module 2

Module 3

\((a+b)x3\)

Enterobactin

Nonlinear (type C)

Module 1

Nonlinear

Module 2

\(a+(a+b)x2+c\)

Vibriobactin

(i)

(ii)

(iii)
present within a cell. A domains are therefore considered to be the primary determinant of substrate selectivity which dictates the primary structure of the product. The reaction catalyzed by A domains involves a two step mechanism and is analogous to that of aminoacyl-tRNA synthetases in ribosomal protein synthesis. Despite the similarity in substrates and reaction mechanism of these two class of enzymes, they are structurally unrelated (Arnez and Moras, 1997). The first step catalyzed by A domains is the activation of the selected substrate as acyl-adenylate using ATP and then in the next step the activated amino acid is transferred as a thioester to the 4’-phosphopantetheine (4’PP) cofactor of the downstream PCP domain as shown in Figure 1.5.

The A domains belong to superfamily of adenylate-forming enzymes that includes firefly luciferases and acyl-coenzyme A (CoA) synthetases. Despite, the sequence variation and diversity of their origin, all these enzymes share homologous structure and contain a set of highly conserved signature motifs. Two crystal structures of A domains have been solved which include the L-phenylalanine activating A domain (PheA) of multidomain protein GrsA of gramicidin S synthetase from Bacillus brevis (Conti et al., 1997) and the stand-alone 2,3-dihydroxybenzoic acid activating domain (DhbE) of bacillibactin synthetase from Bacillus subtilis (May et al., 2002). Other available structures of acyl adenylate superfamily are luciferase (Conti et al., 1996), acetyl-CoA synthetase (Gulick et al., 2003), long chain fatty acyl-CoA ligase (Hisanaga et al., 2004) and 4-coumurate-CoA ligase (Gulick et al., 2004). Interestingly, the structures of firefly luciferase and PheA show striking similarity despite sharing only 16% sequence identity. The A domains of NRPS share sequence identity of 30%-60%, suggesting that all adenylation domains may have a similar
Chapter One: Review of Literature

Figure 1.6: Structure of the L-Phenylalanine activating adenylation domain (PheA) of the gramicidin S synthetase GrsA. Inset shows the substrate binding site of PheA. The substrate Phe, AMP and two conserved residues Asp235 (red) and Lys517 (blue) are shown in thick bonds. The conserved residues Asp235 and Lys517 are involved in electrostatic interaction with amino and carboxyl group of the substrate, respectively. N- and C-terminal subdomains are shown in yellow and orange color, respectively.

structural fold. In general, adenylate-forming enzymes have a unique subdomain structure, a large amino-terminal (~400 amino acids) and a small carboxy-terminal (~100 amino acids) subdomain with very few contacts between them. The subdomains are linked by a highly conserved linker region. The active site for substrate binding is located at the interface of the two subdomains as apparent in the PheA structure which was co-crystallized in the presence of its amino acid substrate Phe and AMP.

1.4.1.1.1 Substrate specificity of A Domains

The co-crystal structure of PheA with its substrates L-phenylalanine and AMP has provided a great deal of insight into the substrate specificity of A domains, by allowing the accurate identification of the L-phenylalanine binding hydrophobic pocket (Figure 1.6). The substrate binding pocket is defined by 10 residues which are within ~5.5 Å of the Phe substrate. These 10 residues correspond to position 235, 236, 239, 278, 299, 301, 322, 330, 311 and 517 of PheA. Two positions 235 and 517 have
conserved Asp and Lys amino acids, which stabilize the substrate amino acid by making electrostatic interactions with the α-amino and α-carboxyl group respectively. The other eight residues interact with the side chain of substrate amino acid and are involved in substrate selection. An empirical correlation between these residues extracted from 145 A domains and their corresponding substrates have been shown in two independent studies (Challis et al., 2000; Stachelhaus et al., 1999). Subsequently, the availability of second crystal structure of an aryl-acid activating stand-alone A domain, DhbE, reinforced the knowledge about A domains taking similar structure and allowed the extension of the correlation to develop the “nonribosomal code” for substrate specificity. The specificity-conferring code could also be recognized in the free-standing eukaryotic NRPS-like enzymes like α-aminoadipate semialdehyde dehydrogenase, α-aminoadipate reductase and Ebony etc. (Di Vincenzo et al., 2005).

Even though, the specificity-code is conserved for a large number of substrates, a level of degeneracy has also been observed in several cases. The same substrate can be activated by A domains with different pocket residues. For example, the predicted selectivity pocket residues for L-proline activating domains RedM, ProB and PltF are different from those first established for domains like module 1 of TycB, which also activates L-proline (Cerdeno et al., 2001; Thomas et al., 2002). Modeling studies with L-proline as substrate clearly demonstrates that the residues from the lower binding pocket do not make contact and therefore are dissimilar between the two domains (Lautru and Challis, 2004) (Figure 1.7). Similar variations in the active site residues could be explained for amino acid substrates with smaller side chains like Gly, Ala, Ser etc. Parallel modeling approaches to understand recognition of L-threonine suggested involvement and importance of only two out of the ten binding residues in defining its selectivity (Ackerley et al., 2003).

In absence of additional A domain structures, molecular modeling and bioinformatics approach might prove helpful in refining models for recognition of different substrates. Such understanding of substrate specificity of A domains will have important implications in predicting the substrates and consequently the final products of newly discovered NRPS genes. Another major application is in rational designing of novel products by modifying the specificity-conferring residues of known biosynthetic NRPS clusters. The scope of altering the substrate preference has been demonstrated by changing the specificity of L-phenylalanine, L-glutamic and L-aspartic acid activating A domains to L-leucine, L-glutamine and L-asparagine.
Figure 1.7: Models of L-Pro-acivating $A$ domains from (a) module 1 of tyrocidine synthetase TycB and (b) the prodiginine synthetase RedM. It is evident from the substrate modeling that the residues from the lower end of binding pocket have lesser role to play in selection of substrate amino acids with smaller side chains. Figure adapted from Lautru and Challis, 2004.

respectively. However, the newly recognized amino acids (substrates) by these $A$ domains were only marginally different from the original. These rational designing of substrate specificity were achieved through mutations in binding residues to match the corresponding code of latter specificities (Eppelmann et al., 2002; Stachelhaus et al., 1999). Recently, a new computational method has been employed using transductive support vector machines (TSVM) to predict the substrate specificity of $A$ domain, using the physico-chemical properties of the residues lining the active site (within 8 Å of the bound substrate) (Rausch et al., 2005).

1.4.1.2 The Peptidyl Carrier Protein

A central role is played by peptidyl carrier protein ($PCP$) or thiolation ($T$) domains in channeling of the intermediates within as well as between the modules. The $PCP$ domain is a small domain consisting of about 80 amino acids and functions as transporter unit for the intermediates (Stachelhaus et al., 1996). Substrates activated by $A$ domain are fixed as thioester on the $PCP$ domain before they are passed on to different catalytic centers. Also the growing polypeptide intermediates remain covalently attached as thioester to its $\sim$20 Å 4'-phosphopantetheine (4'PP) cofactor.
Chapter One: Review of Literature

This cofactor is post-translationally transferred to a conserved Ser residue of apo-PCP by 4'-phosphopantetheinyl transferase (Lambalot et al., 1996), and acts as a flexible swinging arm to take the bound amino acyl or peptidyl substrate from one catalytic domain to the next (Figure 1.2). Very recently such large movement of the 4'PP arm has been shown by nuclear magnetic resonance (NMR) spectroscopy (Koglin et al., 2006), thus providing proof for this "swinging-arm" model. The PCP domains belong to carrier protein (CP) superfamily which includes acyl carrier protein (ACP) of PKS and FAS. Even though, these CPs show a high degree of sequence variation, they have a consensus GxxS (core-T) motif which includes the Ser conserved for cofactor binding. All the structures solved for CPs show strong resemblance. They have an antiparallel four-helix bundle fold with a long loop between first and second helix (Figure 1.8) (Crump et al., 1997; Findlow et al., 2003; Holak et al., 1988; Weber et al., 2000a; Wong et al., 2002; Xu et al., 2001).

1.4.1.2.1 Role of PCP domains in protein-protein interaction

The PCP domain carries the activated substrate and channels it to active site of various functional domains. Many studies have attempted to address the role of CP domain in protein-protein interactions. These studies on interaction of CPs have shown helix-2 as the key element for recognition with its partners (Finking et al., 2004b; Mofid et al., 2002; Parris et al., 2000; Zhang et al., 2001). The ability of helix-3 to adopt multiple conformations as seen in structural studies on PKS frenolicin ACP (Li et al., 2003) and E. coli FAS ACP (Kim and Prestegard, 1989) also suggests helix-3 as an important element in mediating interactions with other partners of the CPs. Recently mutational studies on CP of enterobactin synthase system, identified three residues crucial for its interaction with its downstream C domain. While one of these residue is on helix-2, interestingly, the other two are present on the surface of helix-3 (Lai et al., 2006). It was also shown that these mutations did not affect the interaction of the CP with the A domain, suggesting that different partners probably recognize separate regions of CPs. Thus, available studies on CPs indicate three regions to be important in mediating protein-protein interaction with the partner domains. These regions include first loop (between helix-1 and helix-2), helix-2 and helix-3 of most CPs.
Figure 1.8: Common folds of carrier proteins: (a) Secondary structure representation of (a) PCP from NRPS (PDB code 1DNY); (b) ACP from PKS (PDB code 2AF8); (c) ACP from FAS (PDB code 1T8K). The loop between helix-1 and helix-2 of these carrier proteins is marked with arrows. The conserved cofactor binding residue Ser is shown in stick and its sequence number in respective structures are mentioned. 1DNY and 2AF8 are minimized NMR structures while 1T8k is a crystal structure. (b) Superimposed structures of the three carrier proteins. The residues which are part of helix in the three structures are shown in blue color in the structural alignment, and those which were used to superimpose the structures are boxed. Conserved Ser for all the structures are shown in sticks.
Figure 1.8
Figure 1.9: (a) Reaction mechanism of C domains. The nucleophilic attack of free amino (or imino, hydroxyl) group of the acceptor amino acyl bound to the downstream PCP onto the acyl group of the amino- or peptidyl-PCP of the upstream module leads to the peptide bond formation with the release of a water molecule. (b) Cartoon representation of VibH, the stand alone C domain, perpendicular to the solvent channel. The two subdomain of the pseudodimer is colored differently and the catalytic His is labeled and its bonds are shown as sticks.

1.4.1.3 The Condensation Domain

The condensation (C) domain catalyzes the formation of a peptide bond between two PCP-bound activated substrates present on adjacent modules resulting in elongation of peptide by one residue attached to the PCP domain in the downstream module (Stachelhaus and Marahiel, 1995). The PCP domain downstream to a C domain possesses an aminoacyl moiety as thioester, while the upstream PCP domain contains a peptidyl group. Condensation domains are typically 420 amino acids long and have a conserved motif (H1HxxxxDG; His motif), which is similar to active-site motif found in well studied acyltransferases like chloramphenicol acetyl transferase and dihydrolipoyl transacetylase (Bergendahl et al., 2002). The second His of this motif has been shown to act as catalytic base in these acyltransferases (Kleanthous et al., 1985), while the conserved Asp (underlined in His motif) plays a structural role (Lewendon et al., 1988). By using a two module system from gramicidin synthetase, the site of peptide bond formation was shown to be associated with the C domain and the importance of His motif as part of active site was shown by site directed mutagenesis (Stachelhaus et al., 1998). The C domain catalyzes the nucleophilic attack of free amino (or imino, hydroxyl) group of the acceptor amino acyl bound to the downstream PCP (with respect to C domain) onto the acyl group of the amino- or peptidyl-PCP of the upstream module, as depicted in Figure 1.9a.
Chapter One: Review of Literature

Figure 1.10: Structural similarity between CoA, 4'PP and NAC. The sulphydryl (SH) terminal region (highlighted in pink background) of all the three molecules is equivalent.

According to the multiple-carrier thiotemplate model (Stein et al., 1996), the C domain possesses a site for nucleophile (acceptor site) and a site for the electrophile (donor site). The crystal structure of VibH, a free-standing C domain of vibriobactin biosynthetic cluster from Vibrio cholerae (Keating et al., 2002), was subsequently solved and allowed deeper insight into the functioning of this domain. VibH is a pseudodimer consisting of C- and N-terminal halves, both made of α-β-α sandwich. Interestingly, the crystal structure clearly showed a solvent channel passing through the protein which can accommodate two 20 Å 4'PP moiety (Figure 1.9b). The substrates bound to the 4'PP of two adjacent PCP domains can thus enter from each side of tunnel to reach the active site which is located in the middle of the tunnel at the interface of the two subdomains. Sequence and secondary structure analysis indicated that other internal C domains as well as cyclization (Cy) and epimerization (E) domains are also likely to adopt a CoA-dependent acyltransferase fold similar to VibH (Keating et al., 2002)

1.4.1.3.1 Specificity of C domains

Several molecules like Coenzyme A (CoA) and N-acetylcysteamine (NAC) show structural similarity with that of the 4'PP arm of PCP domain. Indeed the 4'PP arm of PCP is derived from CoA and are transferred to carrier proteins by dedicated
family of enzymes called phosphopantetheiny transferases (Lambalot et al., 1996). Figure 1.10 shows the similarity in the functional groups at the sulphydryl terminal of CoA, NAC and 4'PP because of which they all can be recognized as surrogate substrates for catalytic domains that recognize 4'PP bound substrate. Studies to determine the substrate specificities of C domains using aminoacyl-S-NAC thioesters as small-molecule substrates showed high specificity of C domains towards the side chain and L/D configuration of the acceptor amino acid, while only enantio-selectivity for the donor amino acid. However, condensation of noncognate aminoacyl acceptors is possible, as variant of the main NRPS peptides can sometimes be observed in vivo [tyrocidines A, B, C and D (Ruttenberg and Mach, 1966) or Val7- and Leu7- surfactins (Peypoux et al., 1991)]. Thus, relaxed substrate specificity is observed at the donor site as compared to the acceptor site. However the stereochemistry of the C-terminal amino acid of the peptidyl chain appears to be an important element in the donor substrate recognition (Clugston et al., 2003), as observed for the acceptor site. Similarly, using aminoacyl-CoA derivatives to directly transfer the aminoacyl-phosphopantetheine to PCP domains, thus bypassing the A domains selectivity, it was shown that the first C domain of tyrocidine synthetase does show low specificity at donor site while high substrate selectivity at the acceptor site (Belshaw et al., 1999). The acceptor site was subsequently shown also to be enantio-selective for the L-amino acid, and therefore would bind the L-aminoacyl-PCP immediately after recognition and activation of the amino acid by A domain (Linne and Marahiel, 2000). Thus the acceptor L-aminoacyl-PCP remains bound at the acceptor site and is not available to be processed downstream unless the peptide bond is formed with the upstream donor group. Once peptide bond is formed, the resulting peptidyl product is no longer recognized by the acceptor site of C domain and is released making it available to downstream domains including the optional E domain. Deletion of C domains from an elongation module would therefore lead to immediate downstream processing of the activated L-aminoacyl-PCP and therefore initiation from that module since there would be no upstream C domain to sequester the L-aminoacyl-PCP. This conversion of internal elongation module to initiation module by deleting the upstream cognate C domain was confirmed experimentally (Linne and Marahiel, 2000). This explains various crucial roles of C domain in the timing of epimerization, preventing the misinitiation from an elongation module and also controlling the directionality of
peptide synthesis apart from catalyzing the peptide bond during nonribosomal peptide synthesis.

1.4.1.4 The Thioesterase Domain

The thioesterase (TE) domain is the terminal domain of a NRPS biosynthetic cluster and is implicated in release of the nonribosomal product. It catalyzes either the hydrolysis or the intramolecular cyclization of the peptidyl chain, leading to a variety of structures like linear peptides (vancomycin), cyclic (tyrocidine-A), oligomeric macrolactones (enterobactin) or branched-chain cyclic lipopeptides (surfactin-A and bacitracin-A). The reaction catalyzed by TE domains proceeds through a peptidyl-O-TE intermediate by cleavage of the thioester bond between the peptide and 4'PP cofactor of PCP domain. The next step for release of the product requires a nucleophilic attack either by water molecule leading to hydrolysis or intramolecular nucleophile leading to macrocyclization (Figure 1.11).

![Figure 1.11: Schematic representation of the product release by NRPS TE domains: The peptide chain is depicted as curved line and potential intramolecular nucleophiles are shown. Once the peptide-O-TE (acyl-enzyme) intermediate is formed in the first step of catalysis by TE domains, the nucleophile attack on the ester bond and the peptide product is released. Nucleophilic attack of water leads to hydrolysis, while intramolecular nucleophile leads to cyclized product. Internal amino (blue) or carboxyl (red) nucleophile leads to branched-chain cyclized macrolactam and macrolactone, respectively.](image-url)
Thus, TE domains can be classified as hydrolyzing or cyclizing based on their type of chain release. The cyclization reaction catalyzed by TEs can be classified as macro lactonization or macro lactamization based on the involvement of carboxyl or amino group (the intramolecular nucleophile which attacks to produce the cyclized product), respectively. Furthermore, the cyclizing TEs can also lead to cyclic or branched-chain product, depending upon whether the intramolecular nucleophile is at the terminal end or present at intermediate position. Some TE domains are also involved in oligomerization along with the cyclization of the product as exemplified by iterative type NRPS e.g. gramicidin S, bacillibactin and enterobactin (Shaw-Reid et al., 1999). The mechanism of product oligomerization in such TE is not known.

Similar TE domains are also found in polyketide synthases (PKS) and perform analogous function. Additionally, few biosynthetic clusters also contain a second thioesterase gene, which is present on a separate protein. These TEs have been referred as type II TE (TE-II), while the TE domains integrated within the NRPS and PKS megasynthases are called type I TE (TE-I). In both PKS (Heathcote et al., 2001) as well as NRPS systems (Schwarzer et al., 2002), type II TEs have been implicated in release of substrates which cannot be processed further.

1.4.1.4.1 Substrate specificity of TE domains

The only structure available for NRPS TE domain (Srf-TE) is from surfactin-A biosynthetic cluster from Bacillus subtilis (Bruner et al., 2002b). It showed that TE domains belong to α/β hydrolase family having a catalytic triad defined by residue Ser80, His207 and Asp107. Co-crystallization of Srf-TE with substrate analogue and boronate inhibitor has led to identification of hydrophobic binding pocket for two C-terminal amino acids (D-Leu-6 and L-Leu-7) of the cyclic heptapeptide substrate surfactin-S (Bruner et al., 2002b; Tseng et al., 2002). Biochemical studies have also extensively probed the substrate specificity of cyclizing TE domains in tyrocidine synthetase (Tyc-TE) and surfactin synthetase (Srf-TE) (Kohli et al., 2002a; Kohli et al., 2001; Traeger et al., 2001; Tseng et al., 2002). These studies indicate that TE domains are selective for the side chains of the substrate residues adjacent to the site of nucleophile or the C-terminus of the nonribosomal linear peptide product. Side chain selectivity and enantio-selectivity is observed for residues next to the intramolecular nucleophile (Glu-1 for surfactin and D-Phe-1 for tyrocidine). Though, the binding site for a D-aromatic residue of substrate has been identified in the Tyc-
Chapter One: Review of Literature

TE (Kohli et al., 2002b), no information could be obtained for similar nucleophile binding pocket in Srf-TE. Influence of various leaving group, length of substrate, and nature of cyclizing nucleophile have also been studied for both Tyc-TE and Srf-TE, but the results are quite different for the two TE domains (Sieber et al., 2004). Overall, the Srf-TE appears to be relatively specific for its substrate (Tseng et al., 2002), whereas Tyc-TE has a rather broad substrate specificity (Kohli et al., 2002a). This broad specificity of Tyc-TE has been exploited to enzymatically cyclize libraries of synthetic compounds and screen them for biological activity (Kohli et al., 2002b). Thus, it is difficult to draw a general conclusion regarding the substrate specificity of TE domains although such information is extremely important in engineering biosynthetic strategies.

1.4.2 The Optional Domains

1.4.2.1 The Cyclization Domain

The modules incorporating Ser, Thr or Cys amino acid often have a specialized class of C domains instead of regular C domains at their amino terminus (see module 2 of bacitracin-A biosynthetic cluster shown in Figure 1.3). These Cyclization (Cy) domain substitute C domains and hence more appropriately called as alternate domains. Cy domain catalyzes amide bond formation and cyclization of the side chain nucleophile with the newly formed amides leading to five-membered heterocyclic rings of the thiazoline and oxazoline family (Patel and Walsh, 2001; Suo et al., 1999). The functional role of the Cy domain in condensation and heterocyclic ring formation has been demonstrated in vitro in yersiniabactin synthetase system (Gehring et al., 1998b). However, little is known about the timing and the molecular mechanism of this reaction. The active site His motif (HHxxxxDG) of C domain is replaced by DxxxxD in Cy domains. These domains carry out cyclization along with the condensation, thus having two activities associated with the same domain. The reaction mechanism for Cy domain is depicted in Figure 1.12. Based on secondary structure predictions, it has been proposed that Cy domains take structural fold similar to VibH, which belongs to CoA-dependent acyltransferase superfamily. The presence of thiazoline and oxazoline functional groups serve to increase the structural diversity of the nonribosomal peptides and enables them to function as an efficient metal-chelating (e.g., vibriobactin) or intercalating molecules (e.g., bleomycin A2). Heterocyclic thiazolidine rings can also be incorporated by trans action of other
enzymes, e.g. NADP-dependent reductase (PchG) in pyochelin synthesis (Reimmann et al., 2001), and oxygenase (Ox) domain in epothilone and bleomycin synthesis (Du et al., 2003).

1.4.2.2 The Epimerization Domain

A striking feature of nonribosomal peptides is the presence of D-amino acids. In cyclosporin-A synthetase this is achieved by selective recognition of a D-amino acid directly by an A domain. However, in most NRPS clusters the A domain recognize L-amino acid and after they are incorporated into the growing peptide chain, E domains catalyze the epimerization of the C-terminal L-amino acid of the PCP bound peptide (Linne et al., 2001; Stachelhaus and Walsh, 2000). Based on in vitro experiments, it has been proposed that the E domain provides a mixture of L- and D- isomer and it is the enantio-selectivity of the downstream C domain that selects the D-isomer for elongation of the peptide on NRPS (Linne and Marahiel, 2000; Luo et al., 2002). E domains are also believed to play an important role in the protein-protein interactions in NRPS clusters involving multiple proteins. E domains do not show homology to known epimerases or N-acyl racemases. They represent a novel class of enzymes for this biological activity. Based on sequence analysis, it has been proposed that E domains adopt CoA-dependent acyltransferase fold similar to C domains. E domains share the His motif with C domains and the conserved second His has also been shown to be the catalytic base for E domains (Stachelhaus and Walsh, 2000). Investigations through Ca-H washout experiments by same authors suggest involvement of two bases in the reaction catalyzed by E domains. While the
second His of His motif was confirmed to be one of these, the identity of the second base could not be determined.

1.4.2.3 The Methyltransferase Domain

Many NRPS clusters contain methyltransferase (MT) domain that are responsible for N-methylation of amino acid residues. These modifications not only increase the diversity of the product but also decrease the susceptibility of the peptide product from various proteases. The MT domains of NRPS catalyze the transfer of the methyl group from S-adenosylmethionine (AdoMet) to the $N_a$ of the thioesterified amino acid releasing S-adenosylhomocysteine (AdoHCy) as depicted in Figure 1.13. Cis-acting N-methyltransferase (N-MT) domains are of approximately 400-420 amino acids long and are inserted between core motifs A8 and A9 of A domains, e.g. in thaxtomine and cyclosporine synthetase. N-methylation takes place after covalent attachment of the amino acid on the corresponding PCP domain prior to peptide bond formation (Hacker et al., 2000). O-methyltransferase (O-MT) and C-methyltransferase (C-MT) catalyze methylation of hybrid NRPS/PKS products. C-MT domains incorporate methyl groups in the final product forming branched chain while the O-MT domains are responsible for methylation of hydroxyls or $\beta$-keto groups present in intermediate bound to the carrier protein. For example, myxothiazole biosynthetic cluster (Silakowski et al., 1999), which has both type of O-MT domains, MtaE modifies hydroxyl while MtaF gives rise to the methylation of $\beta$-keto groups.

![Diagram of N-methyltransferase domains in peptide synthetases. The methyl group colored red in S-adenosylmethionine (AdoMet) is transferred to $N_a$ of PCP bound substrate.](image)
producing enol form. While $C-MT$ is present as integrated domain in these multidomain enzymes, $O-MT$ is generally found as stand-alone domain.

### 1.4.3 The Group Transferases

In addition to the above mentioned enzymatic domains there are other classes of enzymes which are involved in modification of the natural products made by NRPS or PKS megasynthase (Figure 1.14). These enzymes are present as separate proteins and act in *trans* to modify the product in contrast to various *cis* acting integrated core and optional domains of NRPS as discussed in earlier sections. Various enzymes that come under this category are glycosyltransferases (GTr) (e.g., chloroeremomycin), acyl transferase (AT) (e.g., mycobactin), aminotransferase (e.g., mycosubtilin) and halogenating enzymes (e.g., syringomycin) etc.
1.5 HYBRID NRPS/PKS

Numerous recent discoveries have highlighted striking similarities in organization and basic biosynthetic machinery of nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS), despite differences in the overall structure and fundamental building blocks of their products. Both NRPS and PKS are multifunctional megasynthases organized into repeated functional units known as modules, each of which is responsible for a particular stage of polyketide or polypeptide elongation. Modular PKS and NRPS proteins harbor as few as 1 to as many as 11 modules, with the order of the individual modules from amino- to carboxy-terminal matching collinearly with the sequential order of elongation steps in the synthesis of their products. With the increasing number of reported NRPS and PKS gene clusters, mixed NRPS/PKS clusters can be identified, combining both the strategies for the creation of further structural diversity in natural products. These hybrid organizations can have varied combinations of the two multifunctional systems. For example, mycosubtilin which is biosynthesized by 7 NRPS modules and 1 PKS module (Duitman et al., 1999), while rapamycin is biosynthesized by 12 PKS modules and 1 NRPS module (Schwecke et al., 1995). The structure of rapamycin and mycosubtilin with their peptide and polyketide groups highlighted are shown in Figure 1.15. Schematic representation of the model proposed for leinamycin (LNM) biosynthesis is depicted in Figure 1.16, showing the modular organization of hybrid NRPS/PKS megasynthases.
Figure 1.16: Proposed model for Leinamycin (LNM) biosynthesis and modular organization of the LNM hybrid NRPS-PKS megasynthetase. A discrete lnmG AT enzyme loads the malonyl CoA extender units to all six PKS modules. The structures in brackets are hypothetical. Color coding indicates the moiety of LNM that is of peptide (blue), polyketide (red), and other (black) origin. ACP, PCP, and AT and its proposed docking domains are shaded in red, blue, and green, respectively, to highlight the assembly-line mechanism for LNM biosynthesis. A green oval denotes an AT docking domain, and a question mark denotes a domain of unknown function. A, adenylation; ACP, acyl carrier protein; AT, acyltransferase; Cy, condensation/cyclization; DH, dehydratase; KR, ketoreductase; KS, ketosynthase; MT, methyltransferase; Ox, oxidation; TE, thioesterase. Figure adapted from Tang et al., 2004.
Figure 1.16
Chapter One: Review of Literature

The hybrid system consists of functional interface between NRPS and PKS. In other words, the organization may be in such way that the product of NRPS is transferred to PKS or vice versa. Thus, the modules at interface could be NRPS/PKS or PKS/NRPS. In NRPS/PKS hybrids, the C-C bond formation is catalyzed by ketosynthase (KS) domains of PKS, while a C-N (peptide bond) is formed in PKS/NRPS hybrids by C domains of NRPS. The reaction mechanism showing the different nucleophiles in the two types of hybrid arrangements is shown in Figure 1.17. The NRPS and PKS have different quaternary structures as revealed by various biochemical experiments like cross-linking, gel filtration, analytical ultracentrifugation etc. It has been proposed that modular PKS take a homodimeric structure (Staunton et al., 1996), while NRPS systems have a monomeric organization (Sieber et al., 2002). Variations in this general structural organization of NRPS system has recently been shown for vibriobactin synthetase, where an inactive condensation domain is conserved because of its role in dimerization of the whole protein (Hillson et al., 2004).
1.6 RATIONAL MANIPULATION OF NRPS SYSTEMS

The modular architecture of NRPSs allows creating hybrid enzymes by altering the arrangement of modules or domains. Such hybrid enzymes can synthesize novel nonribosomal peptides with desired amino acid sequence. Important features of the product which could be modified by rational design are: (a) length, determined by number of modules, (b) choice of starter and extender units, controlled by A and C domain, (c) modifications by optional domains. Important directions towards rational designing have been extensively discussed by Cane and Walsh (Cane et al., 1998). Some strategies which have been used include domain/module swapping, site directed mutations to alter specificity and homologous recombination in heterologous host. The first successful domain swap experiment was done for A-C didomain of seventh module of surfactin synthetase (van Sinderen et al., 1993), albeit with low yield. Subsequently, A-PCP unit substitution has been reported for actinomycin synthetases (Schauwecker et al., 2000). Inter-modal fusion between A and PCP has also been carried out (Doekel and Marahiel, 2000). Combining PCP and a C domain Mootz et al. demonstrated the feasibility of whole module fusion (Mootz and Marahiel, 1997). Though these examples have been partially successful and present a great potential for the rational design of hybrid enzymes, the appropriate choice of fusion site between the domains is crucial for success of such experiments. The scope of altering the substrate preference by point mutations in A domains to match the specificity-conferring code of desired specificity has been demonstrated by changing the specificity of L-phenylalanine, L-glutamic and L-aspartic acid activating A domains to L-leucine, L-glutamine and L-asparagine respectively (Eppelmann et al., 2002; Stachelhaus et al., 1999). These studies indicated that rational alterations of products could be realized by subtle mutations in the enzyme complex without heavily affecting it by domain fusions.

1.7 COMPUTATIONAL STUDIES OF NRPS

Availability of the complete genome sequences for a large number of microorganisms has revealed numerous examples of gene clusters encoding proteins involved in complex natural product biosynthesis. However, metabolic products of majority of them are yet to be characterized. Hence, in silico analysis of these NRPS clusters can give valuable clues about their metabolic products. Knowledge of conserved signature motifs within domains of NRPS has provided the framework to
identify such domains in unknown protein sequences (Konz and Marahiel, 1999). The crystal structure of phenylalanine activating adenylation domain in complex with its substrate *Phe*, has provided insight into the structural basis of substrate recognition (Conti et al., 1997). Based on this crystal structure computational methods have been developed for prediction of substrate of *A* domains using putative binding residues. (Ansari et al., 2004; Challis et al., 2000; Konz and Marahiel, 1999). Similar approach has also been used for predicting specificity of free-standing adenylation domains like α-aminoadipate semialdehyde dehydrogenase, α-aminoadipate reductase and the protein Ebony (Di Vincenzo et al., 2005). Very recently attempt has also been made to predict specificity of *A* domain using support vector machine (SVM) method (Rausch et al., 2005). Apart from these studies on *A* domains, detailed computational analysis has not been carried out for any other catalytic domains of NRPS. In the present thesis, we report systematic *in silico* analysis of various catalytic domains of NRPS proteins.