Introduction
Nonribosomal peptides produced by microorganisms are secondary metabolites having important biological activities. A large number of these peptide products are pharmaceutically valuable compounds like antibiotic, siderophore, cell wall biosynthesis inhibitor, immunosuppressant, antitumor etc. The biosynthesis of nonribosomal peptides is catalyzed by large multifunctional megasynthase proteins called nonribosomal peptide synthetases (NRPSs). NRPSs are multidomain proteins which resemble polyketide synthases (PKSs) and fatty acid synthases (FASs) in terms of their modular arrangements of catalytic domains. Using this coordinated group of catalytic domains, these enzymes perform sequential condensations of proteinogenic as well as nonproteinogenic amino acid monomers to produce diverse peptide products. The catalytic domains of NRPSs have been defined on the basis of their function. The core domains are adenylation domain (A), for selection of starter and extender units, peptidyl carrier protein (PCP) domain with a 4'-phosphopantetheine (4'PP) swinging arm for transferring intermediates from one active site to another and condensation (C) domain responsible for peptide bond formation. Additional optional domains have also been identified which are responsible for the modification of the peptide backbone. These optional domains include cyclization (Cy), epimerization (E), methyltransferase (MT) etc. Upon reaching its full length, the peptide product is released from the protein in linear or cyclized form by a thioesterase (TE) domain which is generally the last functional domain for synthesis of the peptide product and is present at the distal C-terminus of these enzymes. Polypeptide segments connecting all these domains are referred to as 'linkers'. Typically, each module catalyzes addition of one monomeric unit to the growing nonribosomal peptide chain and generally number of monomeric unit in the final metabolic product correlates with the number of modules in the corresponding NRPS protein. However, various deviations from this biosynthetic paradigm have also been observed.

The elucidation of the biosynthetic machinery in NRPS and hybrid NRPS-PKS gene clusters in different organisms has revealed the enormous potential for generating altered secondary metabolites by manipulating the choice of domains and modules. In several cases, biosynthetic machinery of NRPSs has been successfully altered to generate novel NRPS products. However, the reported genetic manipulation experiments have been restricted to few well characterized gene clusters. With rapid increase in number of NRPS and hybrid NRPS-PKS gene clusters in sequence
databases due to availability of genome sequence, systematic bioinformatics analysis of these gene clusters can be carried out to understand the relationship between the organization of domains/modules and chemical structure of the nonribosomal peptide product. Such analysis can help in harnessing their vast potential of combinatorial biosynthesis more effectively for rational design of novel natural products. This would also help in *in silico* identification of nonribosomal peptide products biosynthesized by uncharacterized NRPS gene clusters found in the newly sequenced genomes.

In this thesis, we have carried out a systematic and detailed computational analysis of the sequence and structural features of various experimentally characterized NRPS and hybrid NRPS-PKS proteins with known nonribosomal peptide or hybrid NRPS-PKS product. Based on this analysis, we have attempted to develop predictive rules which can help in correlating sequence information of NRPS proteins to their substrate specificity and hence to the chemical structure of the corresponding metabolic product. These predictive rules for sequence to product relationship can be used for identifying secondary metabolite products of uncharacterized NRPS clusters and providing guidelines for rational design of novel nonribosomal peptides.

Chapter one describes review of the current literature on different types of NRPSs and the function of various core and optional catalytic domains. This chapter also discusses the available X-ray and NMR structures of various NRPS catalytic domains and earlier bioinformatics studies of NRPS proteins.

Chapter two describes the development of a knowledge-based computational protocol for unambiguous identification of catalytic domains present in a given NRPS protein. The protocol for domain identification has been developed based on threading analysis of various catalytic domains present in a training set of 22 NRPS gene clusters. The prediction accuracy of this novel protocol for identification of NRPS domains was benchmarked on a test set of 32 NRPS and hybrid gene clusters. Since the *A* domains are known to be the primary determinant of substrate specificity, this chapter also discusses development of computational approaches for predicting substrate specificity of *A* domains. Apart from knowledge-based approach, feasibility of predicting substrate specificity by energy-based approach has also been investigated.
Chapter three describes development of a profile based approach for identifying functional family of various proteins which adopt CoA-dependent acyltransferase fold. NRPSs consist of at least three domains, including the essential C domains, which take this structural fold. Such profiles have also been shown to predict the positional preferences of C domains in a NRPS gene cluster. Importance of this information in predicting the order of substrate channeling in biosynthetic clusters consisting of multiple ORFs has been discussed. The feasibility of utilization of profiles has also been studied for functional assignment of proteins adopting CoA-dependent acyltransferase fold from whole genome. The results of genome analysis of *M. tuberculosis* have been reported and discussed. Structural modeling of E domains has led to identification of catalytic base which might be responsible for catalysis of epimerization reaction.

In chapter four, we report the analyses of the sequence and structural features of TE and MT domains present in NRPS, PKS and hybrid NRPS/PKS gene clusters. Substrate specificity of these domains has been addressed. Analyses of active site pocket of macrocyclizing TE domains from NRPSs indicate that the residues lining the base of the substrate binding pocket may play crucial role in determining specificity for the type of peptide they would cyclize. Analysis of MT domains has led to development of HMM profiles for identification of MT domains in a NRPS or PKS cluster and categorizing them as C-, O- or N-MT. Such information about these critical domains would help considerably in prediction of the structure of the final product synthesized by these multidomain megasynthases.

Chapter five describes investigations to understand the crosstalk between NRPS and PKS system for the biosynthesis of two important hybrid NRPS/PKS molecules in *M. tuberculosis*. Analysis of protein-protein interactions in biosynthetic pathway of phthiocerol dimycocerosates (PDIM), the surface-exposed lipids unique to the virulent strains of mycobacterium, led to identification of crucial residues involved in channeling of substrates between acyl carrier protein (ACP) of mycocerosic acid synthase and PapAS, an enzyme having a structural fold similar to C domains of NRPS. This chapter also discusses the investigations carried out to analyze the substrate preferences of accessory enzymes involved in tailoring of mycobactin core. Genomic data mining has also been carried out to understand the role of various acyl-CoA dehydrogenase homologues (fadEs) present in the genome of *M. tuberculosis*. 

3
Chapter six describes the studies on protein-protein interactions between various domains of NRPSs for identification of residues involved in inter-domain recognition. $A$-$PCP$ and $C$-$PCP$ interactions have been investigated in detail as these are likely to play a major role in channeling of substrates and control of substrate specificity. The computational results have been reported and discussed in light of various published experiment data.