Published online 5 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W487–W496***
doi:10.1093/nar/gkq340

SBSPKS: structure based sequence analysis of polyketide synthases

Swadha Anand[1], M. V. R. Prasad[1], Gitanjali Yadav[2], Narendra Kumar[1], Jyoti Shehara[1], Md. Zeeshan Ansari[1] and Debasisa Mohanty[1,*]

[1]National Institute of Immunology and [2]National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India

Received January 31, 2010; Revised April 7, 2010; Accepted April 19, 2010

ABSTRACT

Polyketide synthases (PKSs) catalyze biosynthesis of a diverse family of pharmaceutically important secondary metabolites. Bioinformatics analysis of sequence and structural features of PKS proteins plays a crucial role in discovery of new natural products by genome mining, as well as in design of novel secondary metabolites by biosynthetic engineering. The availability of the crystal structures of various PKS catalytic and docking domains, and mammalian fatty acid synthase module prompted us to develop SBSPKS software which consists of three major components. Model_3D_PKS can be used for modeling, visualization and analysis of 3D structure of individual PKS catalytic domains, dimeric structures for complete PKS modules and prediction of substrate specificity. Dock_Dom_Anal identifies the key interacting residue pairs in inter-subunit interfaces based on alignment of inter-polypeptide linker sequences to the docking domain structure. In case of modular PKS with multiple open reading frames (ORFs), it can predict the cognate order of substrate channeling based on combinatorial evaluation of all possible interface contacts. NRPS–PKS provides user friendly tools for identifying various catalytic domains in the sequence of a Type I PKS protein and comparing them with experimentally characterized PKS/NRPS clusters cataloged in the backend databases of SBSPKS. SBSPKS is available at http://www.nii.ac.in/sbspks.html.

INTRODUCTION

Polyketides constitute one of the largest families of small molecule natural products biosynthesized by microbes, fungi and plants as secondary metabolites. These small molecule natural products not only show enormous diversity in their chemical structures but also have a variety of biomedical and pharmaceutical applications. The elucidation of polyketide biosynthetic machinery by pioneering genetic and biochemical studies has revealed that these secondary metabolites are biosynthesized by a class of enzymes called polyketide synthases (PKSs) using an assembly line mechanism, which resembles fatty acid biosynthesis (1–5). During the last decade, the research on PKS biosynthetic pathways has been pursued with two major goals, namely, identification and experimental characterization of new polyketide natural products in various microbial and fungal species (6,7) and production of novel rationally designed natural products by manipulation of known PKS biosynthetic machinery using biosynthetic engineering approach (3,8,9). Bioinformatics analysis of PKS biosynthetic pathways has played a major role in guiding various experimental approaches towards realizing these objectives (10–13).

PKSDB (14) and NRPS–PKS (15) were the first set of web based computational tools which facilitated correlation of sequences of PKS/NRPS megasynthases to the chemical structures of their corresponding secondary metabolite products. The utility of these tools in secondary metabolite biosynthesis research have been discussed in recent reviews (16,17) and have also been acknowledged in several publications from experimental research groups (12,13). Very recently, softwares like ASMPKS (18), ClustScan (19), CLUSEAN (20) and NP.searcher (21) have been developed for discovery of secondary metabolites by genome analysis. Supplementary Table S1 gives a comparative analysis of the various features in different softwares available for analysis of PKS/NRPS biosynthetic pathways. These newly developed softwares have several additional features, like providing an interface for scanning complete genomes for secondary metabolite biosynthetic gene clusters, prediction of stereo specificity of reductive domains and prediction of the linear chemical

*To whom correspondence should be addressed. Tel: +91 11 26703749; Fax: +91 11 26742125; Email: deb@nii.res.in

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

structures of the polyketide or non-ribosomal peptide chains. Even though these computational tools, including PKSDB and NRPS–PKS, use structural information on PKS or NRPS proteins only to a limited extent, 3D structures of AT domains of PKS and A domains of NRPS have provided valuable clues for formulation of prediction rules for correlating sequences of PKS/NRPS proteins to chemical structures of their secondary metabolite products (14,22). The starter/extender specificity code for A domains was originally proposed by Challis *et al.* (23) and Stachelhaus *et al.* (24) based on analysis of the crystal structure of substrate bound adenylation domain from gramicidin synthetase A (GrsA) (25) and has been implemented subsequently in NRPS–PKS software. Recently Rausch *et al.* (26) have developed the NRPSpredictor tool using a machine learning method like transductive support vector machine (TSVM) and a feature vector consisting of 12 different physico-chemical characteristics of the binding pocket residues derived from the same Phe activating adenylation domain from GrsA. Yadav *et al.* (14) have identified putative specificity determining residues (SDRs) of AT domains of PKS proteins based on the structure of an acyltransferase from *Escherichia coli* FAS (27) and used these SDRs to predict substrate specificities of AT domains. These studies demonstrate that structure-based data can be efficiently used for prediction of functional specificities of PKS or NRPS catalytic domains in view of their conserved structural fold. Recently available crystal structures of various catalytic domains (4,28–32) of Type I PKS and almost complete module of mammalian FAS protein (33) provide novel insight into 3D architecture of catalytic domains within a homodimeric PKS module (32,34). As can be seen from Figure 1, the most notable difference is that, the long stretches of amino acids which are typically depicted as unusually long linker regions by conventional sequence based analysis, in fact adopt compact structural domains. The structural fold adopted by DH–ER linker region is tightly packed with catalytic KR domain in the 3D structure even though, in sequence they are separated by ER domain. This suggests that the polyketide biosynthesis is brought about by complex machinery consisting of a tightly coupled network of core catalytic and structural domains and incorporation of such structural information is crucial for design of domain swap experiments for obtaining novel polyketides by rational design approach. A number of recent studies (35–38) have also demonstrated that the 3D structure adopted by inter-polypeptide linker regions (also called docking domains) of modular PKS proteins plays a crucial role in inter-subunit recognition between cognate ORFs in a modular PKS cluster consisting of multiple ORFs. Hence, a structure based analysis of inter-polypeptide linker sequences can help in predicting cognate order of substrate channeling in a modular PKS cluster. Such structural analysis of docking domains as well as modeling of 3D structure of complete PKS modules have provided valuable clues for the recent experimental discovery of a novel 'intermolecular iterative' mechanism involved in biosynthesis of mycoketides in *Mycobacterium tuberculosis* (39).

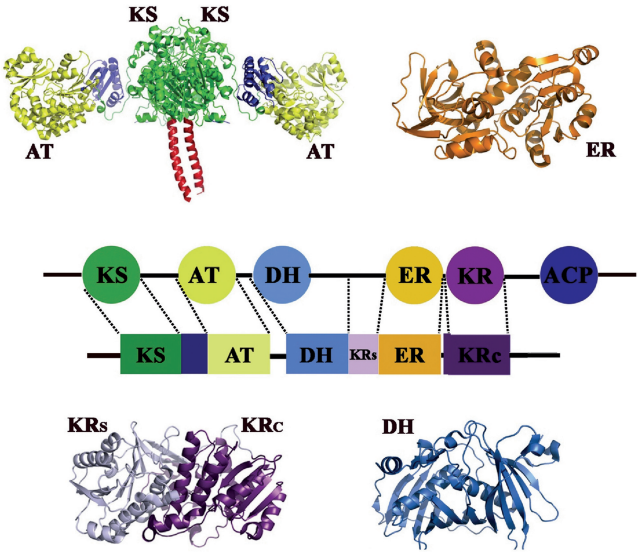

Figure 1. Schematic diagram depicting differences in domain annotations on a typical PKS module by sequence and structure based approach. Various catalytic domains annotated by sequence based programs like NRPS–PKS are shown as circles and the lines connecting the circles represent inter-domain linker stretches. For the same PKS module, the rectangular boxes depict the stretches aligning with various crystal structures of Type I PKS proteins. As can be seen boundaries of some catalytic domains have significantly altered and several amino acid stretches depicted as linkers by NRPS–PKS have compact structures and constitute structural domains of PKS. The figure also shows the crystal structures (PDB IDs 2HG4, 3EL6, 1IZ0, 2FR0 for KS-AT, DH, ER and KR, respectively) of various structural and catalytic domains in the same color as depicted in rectangular boxes.

Therefore, it is necessary to develop web based software tools which can be used to model 3D structures of PKS modules and predict inter-subunit contacts based on the structure of docking domains. Last but not the least, the recent discovery of major deviations from standard PKS paradigms (40) in many organisms as well as characterization of new domains in PKS clusters necessitated the creation of updated databases of experimentally characterized secondary metabolite gene clusters which form the core of sequence based analysis of PKS proteins. This prompted us to develop SBSPKS which not only permits a structure based analysis of PKS proteins, but also permits sequence based comparison of query PKS proteins with a large database of experimentally characterized PKS clusters.

METHODS AND IMPLEMENTATION

The SBSPKS interface is organized in three major functional units, namely MODEL_3D_PKS, DOCK_DOM_ANAL and NRPS–PKS.

Model_3D_PKS

Model_3D_PKS component of SBSPKS has been designed to model the 3D structures of complete modules of Type I modular PKS proteins in biologically active dimeric conformation. Typically the Type I PKS modules consist of four different types of domain

combinations i.e. KS-AT-ACP, KS-AT-KR-ACP, KS-AT-DH-KR-ACP and KS-AT-DH-ER-KR-ACP. However, no crystal structures are available for any of the four different types of complete PKS modules. The only available module crystal structure is KS-AT-DH-ER-KR fragment of mammalian FAS which performs analogous catalytic function and its constituent catalytic domains show structural homology to corresponding PKS catalytic domains. In view of the limited sequence homology between PKS and FAS domains and also the intervening linker regions, it is not possible to model the complete module by straight forward sequence alignment between target (PKS module) and template (FAS module). Model_3D_PKS models individual PKS catalytic domains as well as large inter-domain linkers using available crystal structures of Type I PKS proteins as templates but obtains the relative orientation of various catalytic domains in the PKS module by superposing them onto structurally homologous domains in the mammalian FAS crystal structure. Since mammalian FAS structure lacks the ACP domain, Model_3D_PKS models all PKS modules without the ACP domain. Even though the 3D structure for ACP domain of modular PKS has been elucidated by NMR studies (41), it will not be possible to fix the relative orientation of ACP domain with respect to other catalytic domains in absence of experimental information. In view of the large distances between catalytic centers of various domains in the mammalian FAS structure, it is possible that ACP domain does not have a fixed orientation with respect to other catalytic domains, but is rather mobile to obtain access to various catalytic centers during biosynthesis.

Modeling KS and AT domain along with intervening linker region. The KS-AT fragment of all four types of PKS modules are modeled using the crystal structure (2HG4) (29) of KS-AT di-domain from module 5 of erythromycin PKS cluster as a template. This crystal structure shows that the 100 residue long linker region between KS and AT domains adopts a completely novel compact fold and the relative orientation of the KS and AT domains in the structure 2HG4 is similar to the domain organization seen in mammalian FAS protein.

Modeling catalytic sub-domain of KR. In case of KS-AT-KR-ACP module, the KR catalytic domain as well as the sequences flanking the KR domain are modeled using the crystal structure 2FR0 (30) from module 1 of DEBS. The structure 2FR0 consisting of KR domain along with its flanking linkers (210 amino acids from AT–KR linker and 70 amino acids from KR–ACP linker) has provided novel insight into structural details of KR domain of modular PKS proteins (30). The 210-amino acid AT–KR linker region comprises a new structural sub-domain of KR, while the conventional KR domain along with the 70-amino acid KR–ACP linker forms the catalytic sub-domain. Both sub-domains have identical fold despite lack of significant sequence similarity between them and together they form a tightly packed structure. Similar structural architecture consisting of structural and catalytic sub-domains is also conserved in mammalian FAS. The structure 2FR0 shows significant homology with KR domain as well as flanking AT-KR and KR-ACP linker regions of all KS-AT-KR-ACP modules.

Modeling DH and ER domains. The DH domains of PKS proteins in KS-AT-DH-KR-ACP and KS-AT-DH-ER-KR-ACP modules are modeled using the crystal structure of the dehydratase domain (3EL6) (31) from module 4 of erythromycin PKS. Since no full length ER crystal structures are available for any of the Type I PKS ER domains, the crystal structure 1IZ0, a quinone oxidoreductase structure from *Thermus thermophilus* was used as a template for modeling ER domains.

Modeling structural sub-domain of KR. It has been proposed that DH-KR and DH-ER linkers adopt a fold similar to structural sub-domain of KR and form a compact structure along with catalytic sub-domain of KR. However, in most cases DH-KR and DH-ER linker sequences do not show statistically significant sequence similarity (BLAST *E*-value lower than 10^{-6}) with the structural sub-domain of 2FR0, even though the catalytic sub-domain shows significant similarity. A set of 110 full length KR sequences were collected from 20 representative modular PKS clusters by concatenating DH-KR/ER linker, KR domain and KR-ACP linker sequences. These sequences were grouped into nine representative clusters so that all sequences within a cluster show sufficient similarity with each other over the region corresponding to structural KR, but lack significant similarity across the clusters (Supplementary Figure S1). One representative sequence from each cluster was selected and structural model was built for each of them using 2FR0 as template by threading method. These nine structural models are used as templates for modeling structural sub-domains of KR in cases where DH-KR and DH-ER linkers lacked homology to 2FR0.

Protocol for modeling complete PKS modules. Model_3D_PKS builds homology models for various catalytic and structural domains in any PKS module using the crystal structures 2HG4, 3EL6, 1IZ0 and 2FR0 or any of its nine representative structurally homologous threading based models as templates. The side chain coordinates of these homology models are built using the SCWRL program (42) based on the alignment of the query sequence with the respective templates. The relative orientations of various domains in the 3D structure of the modeled PKS module is fixed by superposing the template structures on the corresponding fragments of the crystal structure of mammalian FAS module. In fact, the template structures are kept pre-aligned on the FAS structure using the DALI server so that Model_3D_PKS does not have to carry out any compute intensive structural superposition while building the model for the complete module. Thus, the modeling protocol of Model_3D_PKS essentially involves aligning the query module sequence to the sequences of the various templates using a local version of the NCBI BLAST

program and based on these alignments, modeling the side chain coordinates on the transformed coordinates of the templates using SCWRL.

Development of Model_3D_PKS web interface. Appropriate interfaces have also been developed for obtaining alignments of query sequence with various templates, downloading the coordinates of individual domains in the module or visualizing them using JMOL applet. Suitable utilities have also been provided in Model_3D_PKS for identifying contacting residues between any two domains in the module, and also depicting SDRs in catalytic sites of AT and KR domains. The prediction of starter/extender substrate specificity of the modeled AT domain is carried out based on 13 SDR motifs proposed by Yadav *et al*. (14). Similarly various motifs identified by Caffrey (43,44) and Keatinge Clay (45) are used to predict the stereo specificity of KR domains for orientation of hydroxyl group and the stereochemistry of the R group substituent at alpha carbon. The program can also identify catalytically inactive KR domains based on the motifs proposed by Keatinge Clay (45).

Dock_Dom_Anal

Dock_Dom_Anal interface of SBSPKS permits evaluation of crucial inter-subunit contacts between two ORFs in a modular PKS gene cluster. The inter-subunit contacts are identified based on the assumption that a single helical stretch from the C-terminus linker of the preceding ORF and three helical stretches from the N-terminus linker of the succeeding ORF together form a four helix bundle structure called 'docking domain'. Structure of a docking domain from erythromycin PKS cluster has been elucidated both by NMR (35) as well as crystallographic (46) studies. It has been proposed that two crucial electrostatic residue pairs in the docking domain structure mediate inter-subunit association during substrate channeling between multiple ORFs in a modular PKS cluster, while unfavorable contacts at equivalent positions in the docking domain are believed to discriminate non-cognate inter-subunit association. This has been referred to, in the literature as 'docking code' (35,36,47). Site directed mutagenesis experiments (48) as well as evolutionary analysis (37,38) of cognate and non-cognate residue pairs in experimentally characterized modular PKS clusters have provided evidence in support of docking code. Recently Yadav *et al*. (37) analyzed N- and C-terminus linker sequences of a large number of modular PKS clusters and have developed a simple scoring scheme based on the 'docking code' by which cognate combinatorial order of substrate channeling can be distinguished from large number of non-cognate combinatorial possibilities. Dock_Dom_Anal interface of SBSPKS has implemented the computational protocol developed by Yadav *et al*. (37) for predicting the preferred order of substrate channeling for modular PKS cluster consisting of multiple ORFs. Here we give a brief description of the protocol, while additional details can be found in the work of Yadav *et al*. (37). Given the N-terminus linker of the preceding ORF and C-terminus linker of the succeeding ORF, Dock_Dom_Anal identifies crucial interacting residue pairs based on structure based sequence alignment with the NMR structure of the docking domain. The two interacting residue pairs on the interface between the two ORFs are categorized as favorable, unfavorable and neutral based on a simple scoring scheme described in the earlier work by Yadav *et al*. (37). Supplementary Figure S2 shows a schematic diagram depicting the protocol used by Dock_Dom_Anal for extracting interface contacts from sequences of inter-polypeptide linkers. Given the N- and C-terminus linker sequences of a series of ORFs from a modular PKS cluster, Dock_Dom_Anal combinatorially evaluates all possible interface contacts and ranks each ORF combination in terms of total number of favorable and neutral interface contacts. Since the first ORFs of modular PKS clusters contain loading modules with different domain combinations and last ORFs typically have thioesterase (TE) domains for release of the polyketide chains, the identity of the first and last ORF can often be inferred based on these features. Dock_Dom_Anal provides option for combinatorial evaluation of interface contacts for the remaining ORFs when the identity of the first and last ORF is provided by the user. This can help in improving the accuracy of the prediction of cognate order of substrate channeling in modular PKS clusters. Dock_Dom_Anal also provides options for extracting the concatenated linker sequences for each of the interfaces in a given combination of ORFs, their alignment with the sequence of the erythromycin docking domain, and crucial interacting residue pairs.

NRPS–PKS

NRPS–PKS interface of SBSPKS server provides a number of user friendly tools for sequence based analysis of putative Types I and III PKS proteins. It can help in identifying and pictorially depicting various catalytic domains present in the sequence of Type I PKS proteins and comparing them with a large number of experimentally characterized PKS and hybrid PKS/NRPS clusters present in the back end databases. NRPS–PKS interface is essentially the latest updated version of the web server developed earlier by our group for sequence based analysis of PKS/NRPS proteins. However, the current version of NRPS–PKS provides access to a much larger data set of experimentally characterized PKS and NRPS gene clusters. NRPS–PKS has been incorporated in SBSPKS server, not only for integrating the various tools for sequence based analysis of PKS megasynthases with the recently developed tools for structure based analysis, but also for accessing the large knowledge base of experimentally characterized PKS clusters and correlating various sequence/structural features of PKS proteins to the chemical structures of their metabolic products. The current database comprises of 167 experimentally characterized PKS and NRPS gene clusters (Supplementary Table S2) consisting of ~4400 PKS and NRPS catalytic domains. The various different types of

catalytic domains present in these 167 gene clusters comprise of 716 KS, 571 AT, 384 DH, 107 ER, 596 KR, 850 PP, 100 TE, 68 MT, 365 C, 354 A, 448 PCP and 23 CHS domains (Supplementary Table S3). Apart from these typical domains present in PKS and NRPS gene clusters, this backend database of SBSPKS also contains few examples of unusual domains like EC (Enoyl CoA hydratase) and PH (Phytanoyl CoA hydroxylase) in curacin (49) and PS (Pyran synthase) in bryostatin (50), which perform novel catalytic reactions and hence, increase structural diversity of secondary metabolites. In most PKS clusters, additional chain modifications are carried out by tailoring enzymes or *trans* PKS enzymes to impart structural diversity and biological activity to the polyketide products. However, these EC, PH and PS domains instead of acting in *trans* were found to be present within the PKS modules along with the catalytic and reductive domains in bryostatin and curacin PKS clusters.

RESULTS

We describe here, a typical usage of SBSPKS for various types of analysis of PKS proteins. Model_3D_PKS and Dock_Dom_Anal interfaces of the SBSPKS software can be used for modeling 3D structure of a PKS module and analyzing inter-subunit interactions in a modular PKS cluster consisting of multiple ORFs. However, the input required for this analysis, i.e. the sequence of a single PKS module and sequences of N- and C-terminus linkers comes from sequence based analysis of PKS proteins. Therefore, a convenient entry point for usage of SBSPKS is the NRPS–PKS interface which is essentially an updated version of the NRPS–PKS software developed earlier by our group for sequence based analysis of PKS and NRPS megasynthases. Figure 2 shows a typical usage of NRPS–PKS interface for identifying various catalytic domains in the query sequence of a type I modular PKS protein and depiction of domains, linkers and modules in a pictorial representation. As can be seen, upon clicking the links on any catalytic domain, the software provides additional interfaces for comparison of the corresponding domain sequence to the homologous domains present in experimentally characterized PKS clusters cataloged in backend databases and alignment with sequences of homologous crystal structures of Type I PKS proteins present in PDB. In case of domains like AT, which are involved in selection of starter and extender substrates, NRPS–PKS extracts the list of residues lining the active site pocket from alignments with homologous PDB structures and attempts to predict the substrate specificity by comparing active site motif of the query domain to the motifs present in AT domains of known specificity (Figure 2). It may be noted that, while most of these analysis are similar to the analysis carried out by earlier version of NRPS–PKS

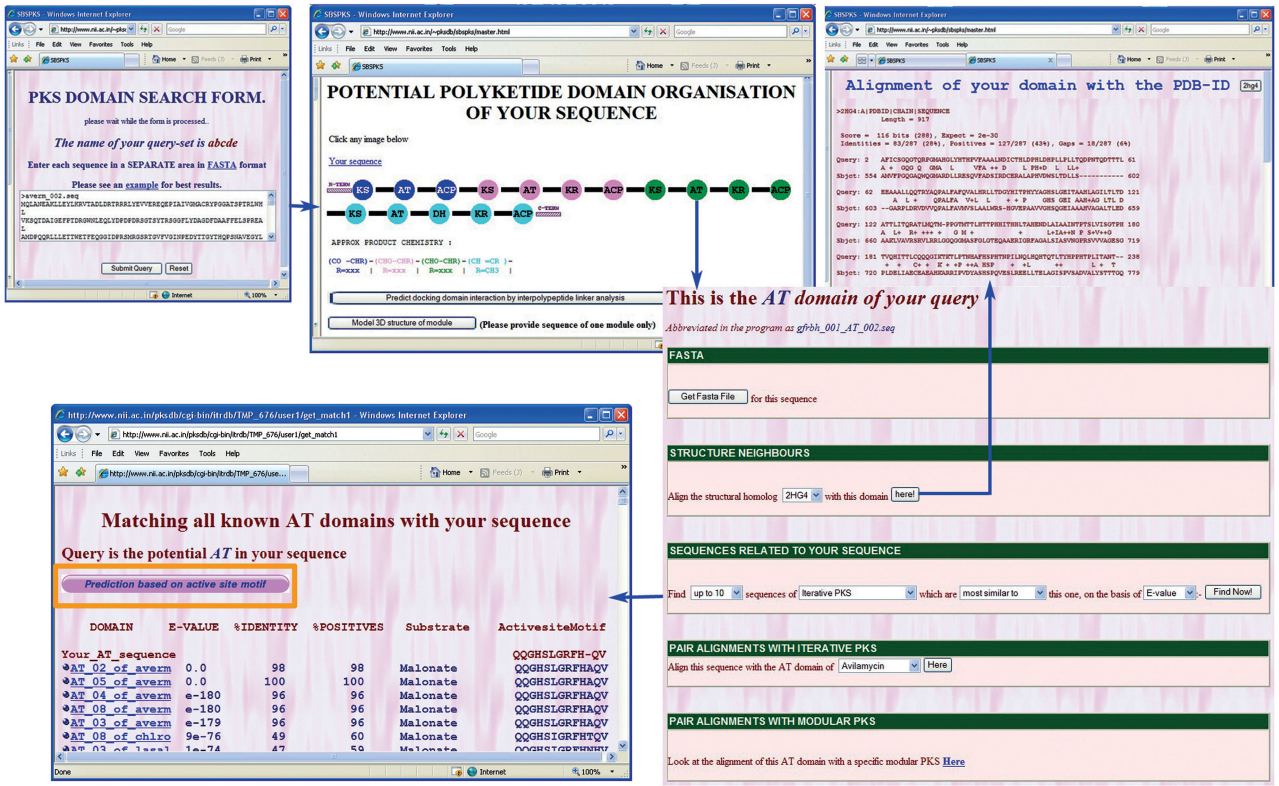

Figure 2. The figure depicts usage of NRPS–PKS for depicting various catalytic domains present in the query sequence. Clicking on each domain leads to a page which provides the details of its alignment with its structural homologs present in PDB as well as with other homologous domains in different experimentally characterized PKS clusters. The screenshot also depicts the prediction of substrate specificity of AT domain based on comparison of its putative active site pocket residues with AT domains having known substrates.

Table 1. Benchmarking of the prediction of AT substrate specificity and comparison with predictions by other softwares

Substrate[a]	Total data set size	Number of correct predictions			
		SBSPKS	ASMPKS	Clustscan	Minowa *et al.* (51)[b]
Malonate training: 136, test = 135	271	132/135			
Methylmalonate training: 107, test = 107	214	106/107			
Ethylmalonate	14	10/14	0/14	8/14	7/10
Methoxymalonate	7	5/7	0/7	1/7	10/12
Propionate	2	2/2	–	–	3/3
Isobutyrate	2	0/2	–	–	2/3
Glycerate	2	1/2	–	–	–
2-Me butyrate	3	2/3	–	–	0/2
Benzoate	2	2/2	–	–	–

[a]Prediction for specificity and sensitivity for Malonate and methylmalonate was carried out by dividing the data set into training and test sets. The values for specificity and sensitivity obtained are Malonate: Sp: 99.1%, Sn: 97.77% and methylmalonate: Sp: 95.86%, Sn: 99.06% The number of correct predictions for other substrates has been calculated by leave-one-out cross-validation approach.
[b]Based on the results reported by Minowa *et al.* (51), the dataset is different from what has been used for benchmarking of SBSPKS and other softwares.

software, the superiority of NRPS–PKS interface of SBSPKS arises from enhancement in number of experimentally characterized PKS clusters available in backend databases. For example, in contrast to less than 200 AT domains of known specificity present in earlier version of NRPS–PKS, backend database of SBSPKS has 571 AT domains with known substrates. Analysis of substrates for these 571 AT domains indicates that, they correspond to 13 different starter/extender substrates. Thus, using the NRPS–PKS interface of SBSPKS web server, substrates can in principle be predicted for 13 substrate specific subfamilies of AT domains, in contrast to SEARCHPKS or earlier version of NRPS–PKS which could essentially identify malonate and methylmalonate specific AT domains. We carried out a benchmarking of the accuracy of NRPS–PKS for prediction of different substrates. In our dataset of 571 AT domains, apart from malonate and methylmalonate, AT domains corresponding to other substrates were relatively fewer in number. Out of the 13 different starter/extender substrates, only those having at least two examples could be considered for the benchmarking analysis (Table 1). Thus, the AT domains specific for unusual substrates like 3-methyl-butyrate, cyclohexanecarboxylic acid, 3-amino-5-hydroxy benzoic acid and *trans*-cyclopentane-(1R, 2R)-dicarboxylic acid were not considered for benchmarking analysis as our data set contained single examples for them. During benchmarking, cross validation for malonate and methylmalonate specific AT domains were carried out by dividing the data into training set and test set, while for other AT domains leave-one-out cross validation approach was used (Table 1). Table 1 lists the total number of AT domains in our dataset for each substrate type, number of AT domains of each type in the test set and the number of correct predictions. As can be seen from Table 1, NRPS–PKS interface of SBSPKS cannot only predict substrates for malonate and methylmalonate specific AT domains with high specificity and sensitivity values, it can also predict other unusual substrates like ethylmalonate, methoxymalonate, propionate, 2-Me butyrate and benzoate, etc with reasonable accuracy. Table 1 also shows prediction results for these unusual substrates by other softwares like ASMPKS and CLUSTSCAN using the same data set. As can be seen, both ASMPKS and CLUSTSCAN can predict for only two other substrates apart from malonate and methylmalonate specific AT domains. The prediction accuracy of SBSPKS for these unusual substrates is in fact comparable to the prediction results reported by Minowa *et al.* (51) using a similar leave-one-out cross validation approach (Table 1). However, it should be noted that, the datasets used are not identical and subtle differences also exist in the prediction protocol. SBSPKS predicts AT specificity using the 13 crucial active site pocket residues identified by Yadav *et al.* and only when multiple hits are found in the data base with identical active site motifs, the results are sorted based on degree of similarity in the entire sequence. On the other hand, Minowa *et al.* (51) use a HMM profile derived from a set of 99 crucial class specific residues of AT domains. Thus our benchmarking analysis demonstrates that SBSPKS can identify AT domains specific for a number of different substrates with reasonably high accuracy.

The other new features of NRPS–PKS are integration of sequence based analysis with new interfaces of SBSPKS for structure based analysis. As can be seen from second panel in Figure 2, the domain depiction page of NRPS–PKS provides links for prediction of docking domain interactions and upon clicking this link, the N- and C-terminus linkers of the sets of ORFs analyzed by NRPS–PKS are automatically provided as input to Dock_Dom_Anal. Similarly the sequence of a module extracted from a Type I PKS protein based on domain boundaries provided by NRPS–PKS can be given as input to Model_3D_PKS for depicting the additional structural domains encoded by the long inter-domain linkers and a series of other structure based analysis. Figure 3 shows typical screen shots for various structure based analysis which can be carried out by providing the

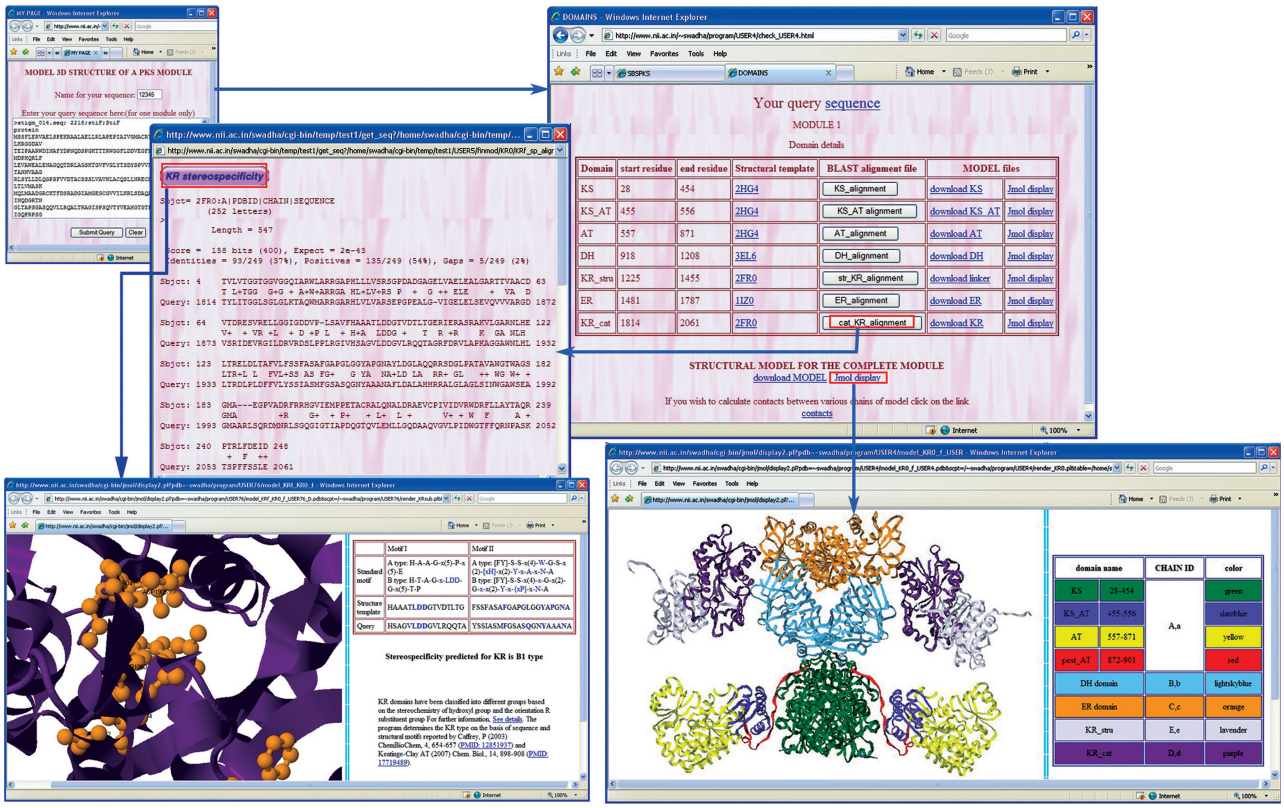

Figure 3. Screen shots showing various options available in Model_3D_PKS and their usage for modeling 3D structures of various structural and catalytic domains present in a PKS module as well as structure of the complete module. The figure also shows usage of the software for depicting SDRs of KR domain on its structure and predicting stereo-specificity of the KR domain based on these residues.

sequence of one PKS module as input to Model_3D_PKS. As can be seen, apart from the catalytic domains, the software also identifies structural domains in the KS-AT and DH-ER linker regions. It also provides links to the structural models of individual catalytic and structural domains present in the module as well as, to the model of the complete module. Upon clicking the button for displaying the 3D structure of the complete module, the dimeric structure of the complete module is shown using JMOL with various structural and catalytic domains depicted in different colors in a new window. All possible manipulations can be done on the displayed model using various features of JMOL. As can be seen from Figure 3, Model_3D_PKS software provides options for viewing alignment of different catalytic domains present in the query module with their respective structural templates and also links to PDB for additional information on these structural templates. The alignment page also provides links for predicting stereo-specificity of KR domains and starter/extender specificity in case of AT domains. As can be seen from Figure 3, upon clicking the stereo-specificity link for catalytic KR domain, the SDRs are depicted on the ribbon model and based on these residues stereo-specificity of KR is predicted. Similar analysis and depiction of SDRs is also possible for AT domain. It may be noted that NRPS–PKS also extracts SDRs for AT domains based on alignment of the query sequence with crystal structures, but display of SDRs by Model_3D_PKS on the structural model can provide additional insight into the structural basis of substrate selection. Model_3D_PKS also permits analysis of inter-domain contacts for selected domain pairs at different cut off distances. Supplementary Figure S3 shows a typical example of calculation of contacts between KS and DH domains and depiction of contacting residues on the 3D structural model. As can be seen, because of the dimeric structure of the module, apart from the contacts between KS and DH domains in the same chain certain inter-chain contacts involving these domains are also present. Analysis of such contacts is crucial for successful design of domain swap experiments for generating novel polyketides.

The other major structure based analysis permitted by SBSPKS is the analysis of inter-subunit contacts in modular PKS clusters consisting of multiple ORFs and use them for predicting the order of substrate channeling. As mentioned earlier, given the FASTA sequences of a set of ORFs as input, NRPS–PKS can automatically extract the N- and C-terminus linkers and inputs those sequences to Dock_Dom_Anal for analysis. Alternatively linker sequences extracted by other programs can also be directly provided in input text boxes of Dock_Dom_Anal. Figure 4 shows a typical example from nanchangmycin modular PKS cluster consisting of

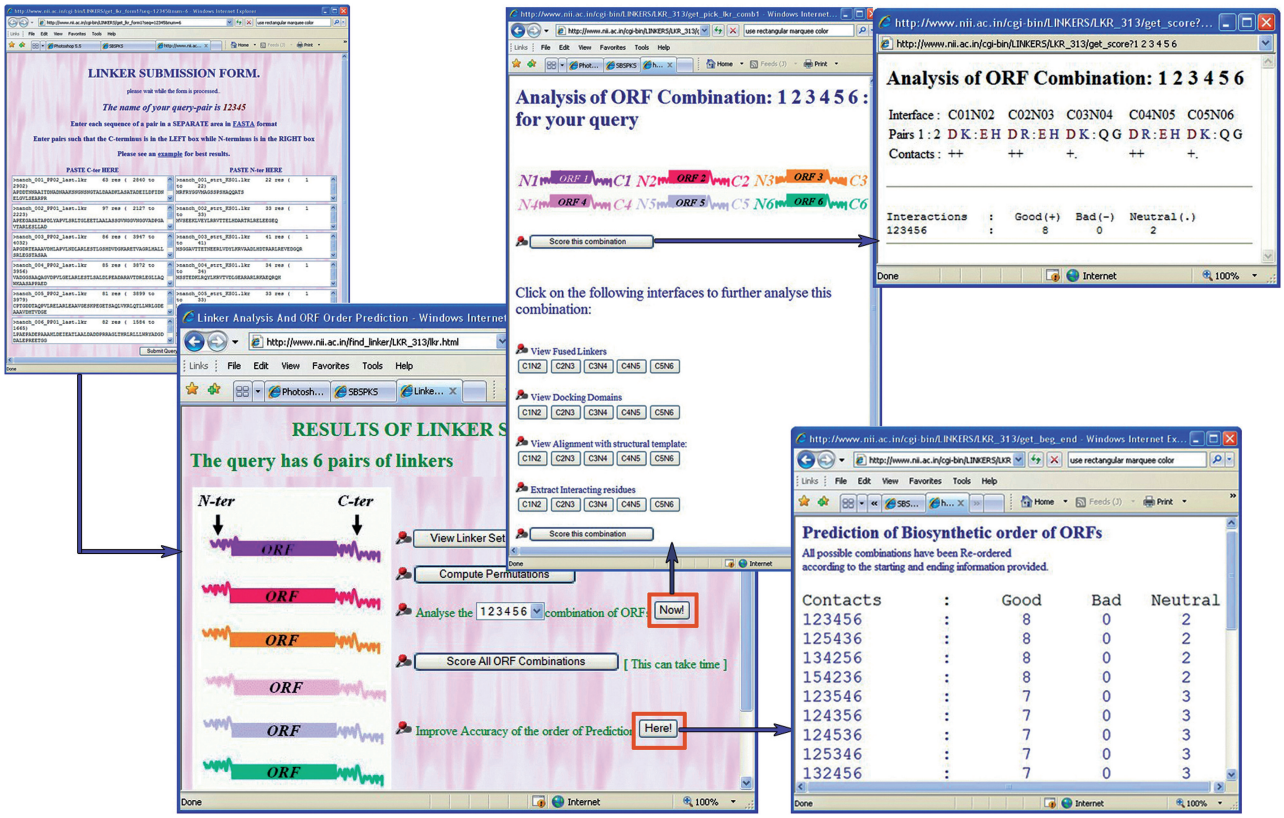

Figure 4. The figure depicts various options available in Dock_Dom_Anal for prediction of the order of substrate channeling between multiple ORFs in a modular PKS cluster based on docking domain interactions. As can be seen the program can quickly evaluate the total number of favorable, unfavorable and neutral inter-subunit contacts for various combinatorial orders of six ORFs present in a modular PKS cluster.

six ORFs. Upon clicking the submit button, the software displays a series of options for viewing the input sequences, analyzing any selected combination of these six ORFs or scoring total number of favorable, unfavorable and neutral contacts in all the five inter-subunit interfaces present in each and every possible combination of these six ORFs. As can be seen, upon selecting a specific combination of these six ORFs for analysis, the software provides further options for aligning the linker sequences to the structural template of the docking domain, extracting crucial interacting residue pairs for each interface and scoring the combination. Dock_Dom_Anal also provides an option for improving the accuracy of cognate order prediction by providing the identity of the first and last ORF from additional knowledge about the gene cluster. In such cases the total number of combinatorial possibilities reduces since first and the last ORFs are kept fixed and this helps in higher scoring of the cognate combination. As can be seen from Figure 4 when ORF1 and ORF6 are designated as first and last ORF, out of 24 different combinatorial arrangements of the remaining four ORFs, the combination corresponding to the cognate order of substrate channeling in nanchangmycin PKS cluster is among the top four combinations with highest score. This demonstrates the utility of Dock_Dom_Anal in deciphering cognate order of substrate channeling by analysis of inter-subunit interactions. Supplementary Table S4 shows the results for benchmarking of Dock_Dom_Anal on a set of 14 modular PKS clusters, with the number of ORFs varying from 4 to 10. For each of these 14 modular PKS clusters the Table S3 lists the total number of combinations (when identity of first and last ORF is fixed, based on information about loading and final chain release domains), total number of favorable, unfavorable and neutral contacts in the cognate combination and the total number of non-cognate combinations having score better, equal to, or lower than cognate combination. As can be seen, for 10 out of these 14 modular PKS clusters, Dock_Dom_Anal can rank the cognate combination within top 20% of the total number of possible combinations. In our earlier work involving the development of computational protocol for prediction of the order of substrate channeling in modular PKS clusters, we had carried out a similar benchmarking on an additional set of 17 modular PKS clusters and in case of 14 out of those 17 clusters the cognate combination could be ranked within top 20%. Thus, the results of our combined benchmarking studies indicate that, for 24 out of 31 modular PKS clusters, valuable clues about the order of substrate channeling can be obtained from analysis of inter-subunit contacts by Dock_dom_Anal interface of SBSPKS. We also analyzed reasons for failure of Dock_dom_Anal in case of the remaining seven modular PKS clusters. It was found that, in many cases due to presence of other catalytic domains, the linker sequences were either too small or too large compared to

typical N- and C-terminus linkers of modular PKS proteins. In some cases, the alignment of linker sequences with helical regions of template docking domains by our completely automated approach was also erroneous due to very low sequence similarity. Therefore, in such cases the program failed to extract correct pairs of inter-subunit contacts. It is also possible that, apart from linkers other catalytic domains also play a role in inter-subunit communication in modular PKS proteins as suggested by some experimental studies. Thus, our benchmarking results suggest that, even though the prediction accuracy of Dock_Dom_Anal is currently limited, it is a valuable tool for preliminary structure based analysis of inter-subunit interactions. Availability of additional sequence/structure data in future can help in further improving its prediction accuracy.

DISCUSSION

SBSPKS provides several interfaces for a variety of sequence as well as structure based analysis of PKSs. The NRPS–PKS interface of SBSPKS can be used for fast automated annotation of organization of catalytic domains, prediction of starter and extender specificity of AT domains and a variety of sequence based comparison with a large number of experimentally characterized PKS clusters having known secondary metabolite products. On the other hand, the Model_3D_PKS interface is useful for identification of structural domains encoded by large linker regions between catalytic domains and modeling interactions between various catalytic and structural domains in a PKS module. Such structural details are extremely important not only for understanding mechanistic details of polyketide biosynthesis, but also for rational design of novel polyketides by biosynthetic engineering approach. Prediction of inter-subunit contacts in modular PKS clusters, based on docking domain analysis using Dock_Dom_Anal can help in identifying cognate order of substrate channeling and deciphering chemical structure of final polyketide products encoded by modular PKS clusters present in genomes. In summary, SBSPKS would be valuable tool for discovery of new secondary metabolite by genome mining as well as rational design of novel polyketides by biosynthetic engineering.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors thank Director, NII for support and Rajesh S. Gokhale for valuable inputs and discussions. S.A. thanks CSIR, India for award of senior research fellowship. The authors are thankful to Prof. R.L. Dunbrack Jr for allowing the use of SCWRL program for modeling of side chains.

FUNDING

Department of Biotechnology (Govt of India) grant (to National Institute of Immunology); BTIS project of Department of Biotechnology grant (to D.M.). Funding for open access charge: Department of Biotechnology grant (to National Institute of Immunology).

Conflict of interest statement. None declared.

REFERENCES

1. Smith,S. and Tsai,S.C. (2007) The type I fatty acid and polyketide synthases: a tale of two megasynthases. *Nat. Prod. Rep.*, **24**, 1041–1072.
2. Hertweck,C. (2009) The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.*, **48**, 4688–4716.
3. Khosla,C., Kapur,S. and Cane,D.E. (2009) Revisiting the modularity of modular polyketide synthases. *Curr. Opin. Chem. Biol.*, **13**, 135–143.
4. Khosla,C. (2009) Structures and mechanisms of polyketide synthases. *J. Org. Chem.*, **74**, 6416–6420.
5. Van Lanen,S.G. and Shen,B. (2008) Advances in polyketide synthase structure and function. *Curr. Opin. Drug Discov. Devel.*, **11**, 186–195.
6. Foerstner,K.U., Doerks,T., Creevey,C.J., Doerks,A. and Bork,P. (2008) A computational screen for type I polyketide synthases in metagenomics shotgun data. *PLoS ONE*, **3**, e3515.
7. Donadio,S., Monciardini,P. and Sosio,M. (2007) Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat. Prod. Rep.*, **24**, 1073–1109.
8. Kittendorf,J.D. and Sherman,D.H. (2006) Developing tools for engineering hybrid polyketide synthetic pathways. *Curr. Opin. Biotechnol.*, **17**, 597–605.
9. Baltz,R.H. (2006) Molecular engineering approaches to peptide, polyketide and other antibiotics. *Nat. Biotechnol.*, **24**, 1533–1540.
10. Zazopoulos,E., Huang,K., Staffa,A., Liu,W., Bachmann,B.O., Nonaka,K., Ahlert,J., Thorson,J.S., Shen,B. and Farnet,C.M. (2003) A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nat. Biotechnol.*, **21**, 187–190.
11. McAlpine,J.B., Bachmann,B.O., Piraee,M., Tremblay,S., Alarco,A.M., Zazopoulos,E. and Farnet,C.M. (2005) Microbial genomics as a guide to drug discovery and structural elucidation: ECO-02301, a novel antifungal agent, as an example. *J. Nat. Prod.*, **68**, 493–496.
12. Simeone,R., Constant,P., Guilhot,C., Daffe,M. and Chalut,C. (2007) Identification of the missing trans-acting enoyl reductase required for phthiocerol dimycocerosate and phenolglycolipid biosynthesis in Mycobacterium tuberculosis. *J. Bacteriol.*, **189**, 4597–4602.
13. Trivedi,O.A., Arora,P., Vats,A., Ansari,M.Z., Tickoo,R., Sridharan,V., Mohanty,D. and Gokhale,R.S. (2005) Dissecting the mechanism and assembly of a complex virulence mycobacterial lipid. *Mol. Cell*, **17**, 631–643.
14. Yadav,G., Gokhale,R.S. and Mohanty,D. (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J. Mol. Biol.*, **328**, 335–363.
15. Ansari,M.Z., Yadav,G., Gokhale,R.S. and Mohanty,D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**, W405–W413.
16. Jenke-Kodama,H. and Dittmann,E. (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. *Nat. Prod. Rep.*, **26**, 874–883.
17. Bachmann,B.O. and Ravel,J. (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.*, **458**, 181–217.
18. Tae,H., Kong,E.B. and Park,K. (2007) ASMPKS: an analysis system for modular polyketide synthases. *BMC Bioinformatics*, **8**, 327.

19. Starcevic,A., Zucko,J., Simunkovic,J., Long,P.F., Cullum,J. and Hranueli,D. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res.*, **36**, 6882–6892.
20. Weber,T., Rausch,C., Lopez,P., Hoof,I., Gaykova,V., Huson,D.H. and Wohlleben,W. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.*, **140**, 13–17.
21. Li,M.H., Ung,P.M., Zajkowski,J., Garneau-Tsodikova,S. and Sherman,D.H. (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.
22. Valenzano,C.R., Lawson,R.J., Chen,A.Y., Khosla,C. and Cane,D.E. (2009) The biochemical basis for stereochemical control in polyketide biosynthesis. *J. Am. Chem. Soc.*, **131**, 18501–18511.
23. Challis,G.L., Ravel,J. and Townsend,C.A. (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, **7**, 211–224.
24. Stachelhaus,T., Mootz,H.D. and Marahiel,M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, **6**, 493–505.
25. Conti,E., Stachelhaus,T., Marahiel,M.A. and Brick,P. (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.*, **16**, 4174–4183.
26. Rausch,C., Weber,T., Kohlbacher,O., Wohlleben,W. and Huson,D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.*, **33**, 5799–5808.
27. Serre,L., Verbree,E.C., Dauter,Z., Stuitje,A.R. and Derewenda,Z.S. (1995) The Escherichia coli malonyl-CoA:acyl carrier protein transacylase at 1.5-A resolution. Crystal structure of a fatty acid synthase component. *J. Biol. Chem.*, **270**, 12961–12964.
28. Khosla,C., Tang,Y., Chen,A.Y., Schnarr,N.A. and Cane,D.E. (2007) Structure and Mechanism of the 6-Deoxyerythronolide B Synthase. *Annu. Rev. Biochem.*, **76**, 195–221.
29. Tang,Y., Chen,A.Y., Kim,C.Y., Cane,D.E. and Khosla,C. (2007) Structural and mechanistic analysis of protein interactions in module 3 of the 6-deoxyerythronolide B synthase. *Chem. Biol.*, **14**, 931–943.
30. Keatinge-Clay,A.T. and Stroud,R.M. (2006) The structure of a ketoreductase determines the organization of the beta-carbon processing enzymes of modular polyketide synthases. *Structure*, **14**, 737–748.
31. Keatinge-Clay,A. (2008) Crystal structure of the erythromycin polyketide synthase dehydratase. *J. Mol. Biol.*, **384**, 941–953.
32. Tsai,S.C. and Ames,B.D. (2009) Structural enzymology of polyketide synthases. *Methods Enzymol.*, **459**, 17–47.
33. Maier,T., Leibundgut,M. and Ban,N. (2008) The crystal structure of a mammalian fatty acid synthase. *Science*, **321**, 1315–1322.
34. Gokhale,R.S., Sankaranarayanan,R. and Mohanty,D. (2007) Versatility of polyketide synthases in generating metabolic diversity. *Curr. Opin. Struct. Biol.*, **17**, 736–743.
35. Broadhurst,R.W., Nietlispach,D., Wheatcroft,M.P., Leadlay,P.F. and Weissman,K.J. (2003) The structure of docking domains in modular polyketide synthases. *Chem. Biol.*, **10**, 723–731.
36. Weissman,K.J. and Muller,R. (2008) Protein-protein interactions in multienzyme megasynthetases. *Chembiochem.*, **9**, 826–848.
37. Yadav,G., Gokhale,R.S. and Mohanty,D. (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol.*, **5**, e1000351.
38. Thattai,M., Burak,Y. and Shraiman,B.I. (2007) The origins of specificity in polyketide synthase protein interactions. *PLoS Comput. Biol.*, **3**, 1827–1835.
39. Chopra,T., Banerjee,S., Gupta,S., Yadav,G., Anand,S., Surolia,A., Roy,R.P., Mohanty,D. and Gokhale,R.S. (2008) Novel intermolecular iterative mechanism for biosynthesis of mycoketide catalyzed by a bimodular polyketide synthase. *PLoS Biol.*, **6**, e163.
40. Wenzel,S.C. and Muller,R. (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. *Curr. Opin. Chem. Biol.*, **9**, 447–458.
41. Alekseyev,V.Y., Liu,C.W., Cane,D.E., Puglisi,J.D. and Khosla,C. (2007) Solution structure and proposed domain domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein Sci.*, **16**, 2093–2107.
42. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L. Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.
43. Baerga-Ortiz,A., Popovic,B., Siskos,A.P., O'Hare,H.M., Spiteller,D., Williams,M.G., Campillo,N., Spencer,J.B. and Leadlay,P.F. (2006) Directed mutagenesis alters the stereochemistry of catalysis by isolated ketoreductase domains from the erythromycin polyketide synthase. *Chem. Biol.*, **13**, 277–285.
44. Caffrey,P. (2003) Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. *Chembiochem.*, **4**, 654–657.
45. Keatinge-Clay,A.T. (2007) A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chem. Biol.*, **14**, 898–908.
46. Buchholz,T.J., Geders,T.W., Bartley,F.E. 3rd, Reynolds,K.A., Smith,J.L. and Sherman,D.H. (2009) Structural basis for binding specificity between subclasses of modular polyketide synthase docking domains. *ACS Chem. Biol.*, **4**, 41–52.
47. Weissman,K.J. (2006) The structural basis for docking in modular polyketide biosynthesis. *Chembiochem*, **7**, 485–494.
48. Weissman,K.J. (2006) Single amino acid substitutions alter the efficiency of docking in modular polyketide biosynthesis. *Chembiochem.*, **7**, 1334–1342.
49. Chang,Z., Sitachitta,N., Rossi,J.V., Roberts,M.A., Flatt,P.M., Jia,J., Sherman,D.H. and Gerwick,W.H. (2004) Biosynthetic pathway and gene cluster analysis of curacin A, an antitubulin natural product from the tropical marine cyanobacterium Lyngbya majuscula. *J. Nat. Prod.*, **67**, 1356–1367.
50. Sudek,S., Lopanik,N.B., Waggoner,L.E., Hildebrand,M., Anderson,C., Liu,H., Patel,A., Sherman,D.H. and Haygood,M.G. (2007) Identification of the putative bryostatin polyketide synthase gene cluster from "Candidatus Endobugula sertula", the uncultivated microbial symbiont of the marine bryozoan Bugula neritina. *J. Nat. Prod.*, **70**, 67–74.
51. Minowa,Y., Araki,M. and Kanehisa,M. (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.*, **368**, 1500–1517.