Towards Prediction of Metabolic Products of Polyketide Synthases: An *In Silico* Analysis

Gitanjali Yadav[¤], **Rajesh S. Gokhale**, **Debasisa Mohanty***

National Institute of Immunology, New Delhi, India

Abstract

Sequence data arising from an increasing number of partial and complete genome projects is revealing the presence of the polyketide synthase (PKS) family of genes not only in microbes and fungi but also in plants and other eukaryotes. PKSs are huge multifunctional megasynthases that use a variety of biosynthetic paradigms to generate enormously diverse arrays of polyketide products that posses several pharmaceutically important properties. The remarkable conservation of these gene clusters across organisms offers abundant scope for obtaining novel insights into PKS biosynthetic code by computational analysis. We have carried out a comprehensive *in silico* analysis of modular and iterative gene clusters to test whether chemical structures of the secondary metabolites can be predicted from PKS protein sequences. Here, we report the success of our method and demonstrate the feasibility of deciphering the putative metabolic products of uncharacterized PKS clusters found in newly sequenced genomes. Profile Hidden Markov Model analysis has revealed distinct sequence features that can distinguish modular PKS proteins from their iterative counterparts. For iterative PKS proteins, structural models of iterative ketosynthase (KS) domains have revealed novel correlations between the size of the polyketide products and volume of the active site pocket. Furthermore, we have identified key residues in the substrate binding pocket that control the number of chain extensions in iterative PKSs. For modular PKS proteins, we describe for the first time an automated method based on crucial intermolecular contacts that can distinguish the correct biosynthetic order of substrate channeling from a large number of non-cognate combinatorial possibilities. Taken together, our *in silico* analysis provides valuable clues for formulating rules for predicting polyketide products of iterative as well as modular PKS clusters. These results have promising potential for discovery of novel natural products by genome mining and rational design of novel natural products.

Citation: Yadav G, Gokhale RS, Mohanty D (2009) Towards Prediction of Metabolic Products of Polyketide Synthases: An *In Silico* Analysis. PLoS Comput Biol 5(4): e1000351. doi:10.1371/journal.pcbi.1000351

Editor: Rama Ranganathan, UT Southwestern Medical Center, United States of America

Received May 20, 2008; **Accepted** March 9, 2009; **Published** April 10, 2009

Funding: The work has been supported by grants to the National Institute of Immunology from Department of Biotechnology (DBT), Government of India, and research grants to DM from BTIS project of DBT, India. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: deb@nii.res.in

¤ Current address: National Institute of Plant Genome Research, New Delhi, India

Introduction

It is well known that polyketide synthase (PKS) gene clusters can generate enormously diverse array of polyketide products by making use of various biosynthetic paradigms like, modular organization of sets of catalytic domains or iterative catalysis of condensation steps using single set of catalytic domains [1]. In view of the pharmaceutical importance of polyketides, there is tremendous interest in identifying PKS gene clusters capable of producing novel polyketides by genome mining. However, the relating the sequence of the various catalytic domains present in a PKS biosynthetic cluster to the chemical structure of the final metabolic product is a major challenge. The availability of the sequences of a large number of experimentally characterized PKS clusters and 3D structural information on homologous protein domains presents a unique opportunity to carry out *in silico* analysis for addressing structural and mechanistic issues concerning polyketide biosynthesis. A number of recent theoretical studies have demonstrated the utility of *in silico* analysis in providing novel insights into the mechanistic details of polyketide biosynthesis as well as in identifying novel natural products by genome mining. Computational analysis of polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) proteins have provided valuable clues for development of knowledge-based methods for identification of catalytic domains in PKS [2,3] and NRPS [4] proteins, prediction of the substrate specificity for AT domains [2,3,5] and adenylation domains [4,6,7]. Such predictions have also been experimentally validated by the recent successful reprogramming of the phthiocerol dimycocerosate (PDIM) biosynthetic pathway in *Mycobacterium tuberculosis* [8] and experimental characterization of a novel exogenous standalone enoyl reductase (ER) involved in PDIM biosynthesis [9]. Bioinformatics analysis of secondary metabolite biosynthetic pathways have also played a crucial role in discovery of novel natural products by genome mining [10–14]. Very recently it has also been demonstrated that, computational analysis of KS domains from trans-AT PKS clusters can give novel clues about the chemical structures of the final polyketide product [15]. Similarly, bioinformatics analysis of docking domain sequences (the original term applied to these regions was "interpolypeptide linker", but the term docking domain is being increasingly used in recent literature) have given novel insight into the evolution of specificity in inter polypeptide interactions in modular PKSs [16]. Pioneering work at Ecopia BioScience using data mining approaches has also

Author Summary

Polyketide synthases (PKSs) form a large family of multifunctional proteins involved in the biosynthesis of diverse classes of therapeutically important natural products. These enzymes biosynthesize natural products with enormous diversity in chemical structures by combinatorial use of a limited number of catalytic domains. Therefore, deciphering the rules for relating the amino acid sequence of these domains to the chemical structure of the polyketide product remains a major challenge. We have carried out bioinformatics analysis of a large number of PKS clusters with known metabolic products to correlate the chemical structures of these metabolites to the sequence and structural features of the PKS proteins. The remarkable conservation observed in the PKS sequences across organisms, combined with unique structural features in their active sites and contact surfaces, allowed us to formulate a comprehensive set of predictive rules for deciphering metabolic products of uncharacterized PKS clusters. Our work thus represents a major milestone in natural product research, demonstrating the feasibility of discovering novel metabolites by *in silico* genome mining. These results also have interesting implications for rational design of novel natural products using a biosynthetic engineering approach.

led to development of proprietary databases which can aid in genomics driven discovery of cryptic biosynthetic pathways [17] and utility of these databases have been demonstrated by identification of novel secondary metabolites [18].

Thus, these studies have established that knowledge based computational approaches can play a powerful role in elucidation of novel secondary metabolite biosynthetic pathways. However, for *in silico* identification of polyketide products of uncharacterized PKS clusters, the computational method should also take into consideration various different paradigms employed by PKS biosynthetic machinery [19]. Several excellent reviews [20,21] describe the type I, type II and type III biosynthetic paradigms. Type I modular PKSs harbor distinct sets of catalytic domains, each set termed as a "module". Each module is responsible for one condensation step and the number of modules in a modular PKS correlate directly with the number of ketide units in its biosynthetic product. In contrast, type I iterative PKSs are characterized by a single set of catalytic active sites which are used iteratively for several rounds of successive condensations till the final product is released. It was initially believed that bacterial PKSs are modular while fungal PKSs function in an iterative manner. However, discovery of mixed PKS clusters involving programmed iterative modules and several other deviations [22,23] from conventional textbook PKS biosynthetic paradigms in various microbes indicate that PKS proteins are not amenable to simple classification based on species of their origin. Therefore, *in silico* methods should be capable of predicting from sequence information, whether a given PKS cluster is iterative, the number of iterative chain condensation steps catalyzed by it and crucial amino acids which control the number of iterations.

In contrast to type I iterative PKSs where a single multifunctional enzyme is involved in biosynthesis of the polyketide product, biosynthesis in type I modular PKS clusters often involve multiple ORFs, each containing several modules. Therefore, predicting the correct order of substrate channeling between various ORFs is crucial for deciphering the final metabolic product of a modular PKS cluster. Several lines of experimental evidence reveal that inter subunit interactions between C-terminal docking domain region of the upstream ORF and N-terminal docking domain region of the downstream ORF, play a crucial role in channeling of substrates from upstream domains to downstream domains [24–27]. Moreover, these interactions involving C-terminus and N-terminus amino acid stretches have been reported to increase the maximum velocity (k_{cat}) of chain transfer of otherwise disfavored substrates by as much as 100-fold [28]. Structural studies using NMR suggest that, these terminal docking domain regions of PKS proteins adopt a specific 3-dimensional fold consisting of a four helix bundle structure [29]. In fact, after the elucidation of this NMR structure, the term 'docking domain' is being increasingly used in the recent literature to describe these terminal amino acid stretches, which were earlier called 'inter polypeptide linkers'. Based on this structure, it has been proposed that recognition between upstream and downstream ORFs in a modular cluster is governed by formation of specific contacts in the docking domain. Several recent experimental studies [30,31] have further validated the role of specific inter polypeptide contacts in controlling inter subunit communication in modular PKS clusters. Very recently NMR studies [32] have also elucidated the role of similar docking domains in governing protein-protein interactions in hybrid megasynthases. Even though these experimental studies have identified specific residue pairs involved in inter subunit recognition, no systematic analysis of experimentally characterized modular PKS clusters have been characterized to investigate whether correct order of substrate channeling in type I modular PKS clusters can be predicted based on these specific inter polypeptide contacts. It may be noted that, even though recent study by Thattai *et al* [16] has attempted to address this question, their algorithm for prediction of PKS multiprotein chain order has been tested on a hypothetical five ORF cluster with only six combinatorial possibilities.

In this work, we have carried out a detailed comparative analysis of the experimentally characterized modular and iterative PKS clusters with known polyketide products to address following major questions relating to *in silico* prediction of polyketide products. Is it possible to distinguish between modular and iterative PKS from their sequence alone? Can we predict the number of iterations a given iterative PKS protein would catalyze and identify crucial amino acid residues that control the number of iterations? Is it possible to predict the correct order of substrate channeling between various ORFs in a modular PKS cluster? We have carried out profile Hidden Markov Model (HMM) analysis of KS domains to identify signature profiles which can decipher whether a PKS protein is modular or iterative. Structural modeling of KS domains of iterative PKS proteins and analysis of their active site pockets have given novel insight into the structural features that dictate the number of iterations catalyzed by a PKS protein and crucial amino acids which control them. Similarly, comparative analysis of crucial inter polypeptide contacts between cognate and non-cognate pairs of ORFs based on the three dimensional structure of the docking domains have given novel clues for prediction of the correct order of substrate channeling.

Results

Distinguishing between modular and iterative PKSs

KS domains are the most conserved among various catalytic PKS domains and are responsible of catalysis of the chain condensation step. We have analyzed them in detail to identify class specific conserved patterns which distinguish modular and iterative PKS systems. For KS domains, the total dataset comprised of 217 pure modular KS domains, 82 pure iterative domains, 19 enediyne, 43 trans-type and 34 KS domains from hybrid NRPS-PKS clusters. Apart from the sequences of 20 experimentally characterized bacterial type I modular clusters included in our earlier analysis [2], an additional set of 18 modular PKS clusters was used as described in Methods. Despite sharing a significant degree of homology ranging from 24% to 40% sequence identity, KS domain counterparts from modular and iterative PKSs and other PKS subfamilies, segregate into distinct clusters in a phylogenetic dendrogram (Figure S1). We have used profile Hidden Markov Models (HMMs) to quantify subtle position specific differences in the probability of occurrence of amino acids in various subfamilies of KS domains (See **methods** for description of various subfamilies). The available KS data set was divided into training and test set, and sequences belonging to the training set were used for building profile Hidden Markov Models by the HMMER package [33]. Benchmarking on the test set indicated that, these HMM profiles were highly sensitive, with a prediction accuracy of 100% for both enediyne and trans-AT sub families, 97% for pure iterative PKSs, 92% for modular KS domains and 88% for hybrid clusters. Therefore, using HMM profiles it is not only possible to distinguish between modular and iterative PKS with a very high accuracy, these profiles can also be used to classify an uncharacterized sequence of a KS domain into various subfamilies within modular and iterative systems. This result has interesting implications for genome sequencing efforts towards identification of novel PKS clusters, because from KS sequence alone, one can get clues about PKS family and decide whether to sequence the entire cluster or not.

Identification of sequence and structural features that control number of iterations

The polyketide products of various iterative PKS proteins are biosynthesized by different number of iterative condensation steps and undergo varying degrees of reductions. Phylogenetic analyses of iterative KS domains revealed that the clustering of iterative PKS sequences is highly correlated with the number of iterations they perform and degree of reductions undergone by the metabolite during biosynthesis (Figure 1). The biosynthesis of polyketides, lovastatin and bikaverin involve eight condensation steps, but their final structures are different because of the different cyclization patterns. Our analysis suggests that, the sequence of KS domain encodes information about chemical structure of the polyketide product. Hence, KS sequences of lovastatin and bikaverin form two different clusters. Based on similar phylogenetic analysis, earlier reports have proposed that KS domains cluster into groups depending on whether the corresponding type I iterative PKS contains additional reductive domains [34–36]. We attribute this feature to a complex programming within the KS domains which enables specific molecular recognition of the products. The observed clustering in Figure 1 could thus be arising from sequence features, that control recognition of specific substrates which have undergone different degrees of chemical and structural modifications due to the presence of reductive domains. Therefore, we wanted to analyze the structural models of various iterative KS domains for identification of specific amino acids or sequence stretches that can potentially control substrate size and extent of unsaturation. The various iterative KS domains were modeled using comparative modeling approach (see Methods for details). The structural templates for various iterative KS domains were identified by BLAST search against PDB or by using threading approach. The *E. coli* KAS-II protein (pdbids 1KAS, 1B3N) were used as the templates for modeling these iterative KS domains. Since 1B3N was a ligand bound structure (Figure 2A), the putative active site pockets (Figure 2B) of various iterative KS structural models could be identified based on amino acids which were in contact with the bound ligand in 1B3N. The structural features of the active site pockets of different iterative KS domains were analyzed further to identify the cavity lining residues (CLRs) and cavity volumes following protocols described in the methods section. Active site residue patterns (Figure 2B) in these structural models allowed us to correlate the cavity volume and hydrophobicity of the active site pockets to the number of iterations and the degree of unsaturation of the polyketide products they synthesize.

The substrate binding cavity in the 1KAS is highly hydrophobic owing to its completely saturated substrate. Polyketides, on the other hand, may contain several hydroxyl groups and unsaturated double bonds. Accordingly, the catalytic pockets in the structural models of polyketide KS domains were found to be less hydrophobic compared to the FAS cavities. Table 1 compares PKS product characteristics with a variety of cavity features. We observed a distinct difference in pocket hydrophobicity within polyketides and it correlated negatively with the extent of unsaturation seen in the product (Figure 3A). For example, the T-toxin PKS model cavity is more hydrophobic than the methylsalicylic acid synthase (MSAS) model cavity and this correlates with the fact that T-toxin is a reducing PKS having a greater proportion of saturated carbons in its final product than the partially reducing MSAS polyketide. Interestingly, cavity volumes correlate positively with the number of iterations (or corresponding product size). We found that polyketide KS cavity volumes fall into three distinct groups; small, large and intermediate (Figure 3B and 3C). The smallest cavities ($\sim$300Å^3) belong to the MSAS type PKSs that perform three iterations. Intermediate sized cavities ($\sim$800Å^3) belong to the napthopyrone (NAP) like PKSs that iterate from five to eight times. The largest cavities, 1780Å^3, were observed for the T-Toxin models that perform 20 iterations. Figure 2B depicts the residues that line the hydrophobic cavity of the template KAS-II protein (volume 934 Å^3) and surround the ligand analogue cerulenin. A comparison of the modeled structures with the template FAS KS structure revealed that in case of MSAS and NAP, the backbones of the models had not altered significantly during modeling (Figure S2), and thus, their functional difference could be traced to specific cavity lining residues (CLRs) (Figure 4). Figure 5A and 5B show the surface topology of the small and intermediate sized cavities. Figure 5A depicts the modeled MSAS KS domain with two tyrosines protruding into the KS cavity from opposite walls and thus blocking the downward flow of the cavity along the dimer interface. These two cavity blocking residues correspond to positions 229 and 400 (1KAS numbering). Interestingly, the conservation profiles of the CLRs shown in Figure 4 revealed that these two *Tyr* residues are highly conserved in all PKSs which carry out three iterations. This further substantiates the important role attributed to these residues based on our structural modeling of the active site pocket. Remarkably, NAP type KS domains have an *Ala* at position 400, that allows the cavity to extend further down thus making their cavities similar to the FAS catalytic cavity, shown for reference in Figure 5C.

Structural analysis thus revealed how substrate binding sites of varying size and hydrophobicity can be generated in type I iterative KS domains by subtle variations of residues on similar backbone folds. The crystal structure of KS-CLF also highlights how specific residues can regulate chain length in type-II PKSs [37]. Our results on role of cavity volume in controlling number of iterative condensations or chain length of type I iterative PKS products is also supported by recent experimental studies involving swapping of KS domains in fungal iterative PKSs, where replacement of

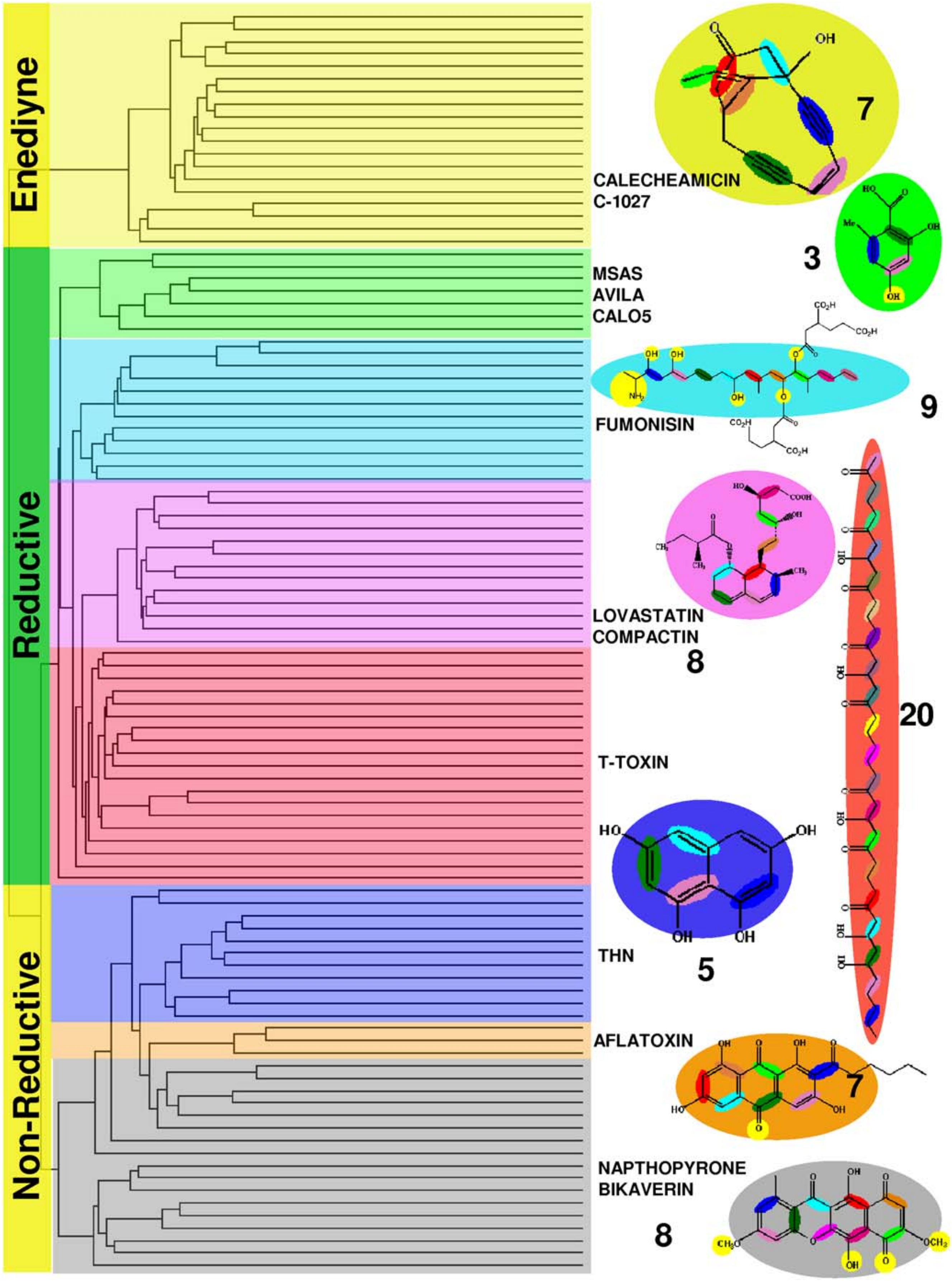


Figure 1. Dendrogram of KS domains from type-I iterative PKS clusters. The branches of the dendrogram have been colored according to the number of iterations catalyzed by the corresponding KS domain. The corresponding polyketide structures have been depicted in the same color.
doi:10.1371/journal.pcbi.1000351.g001

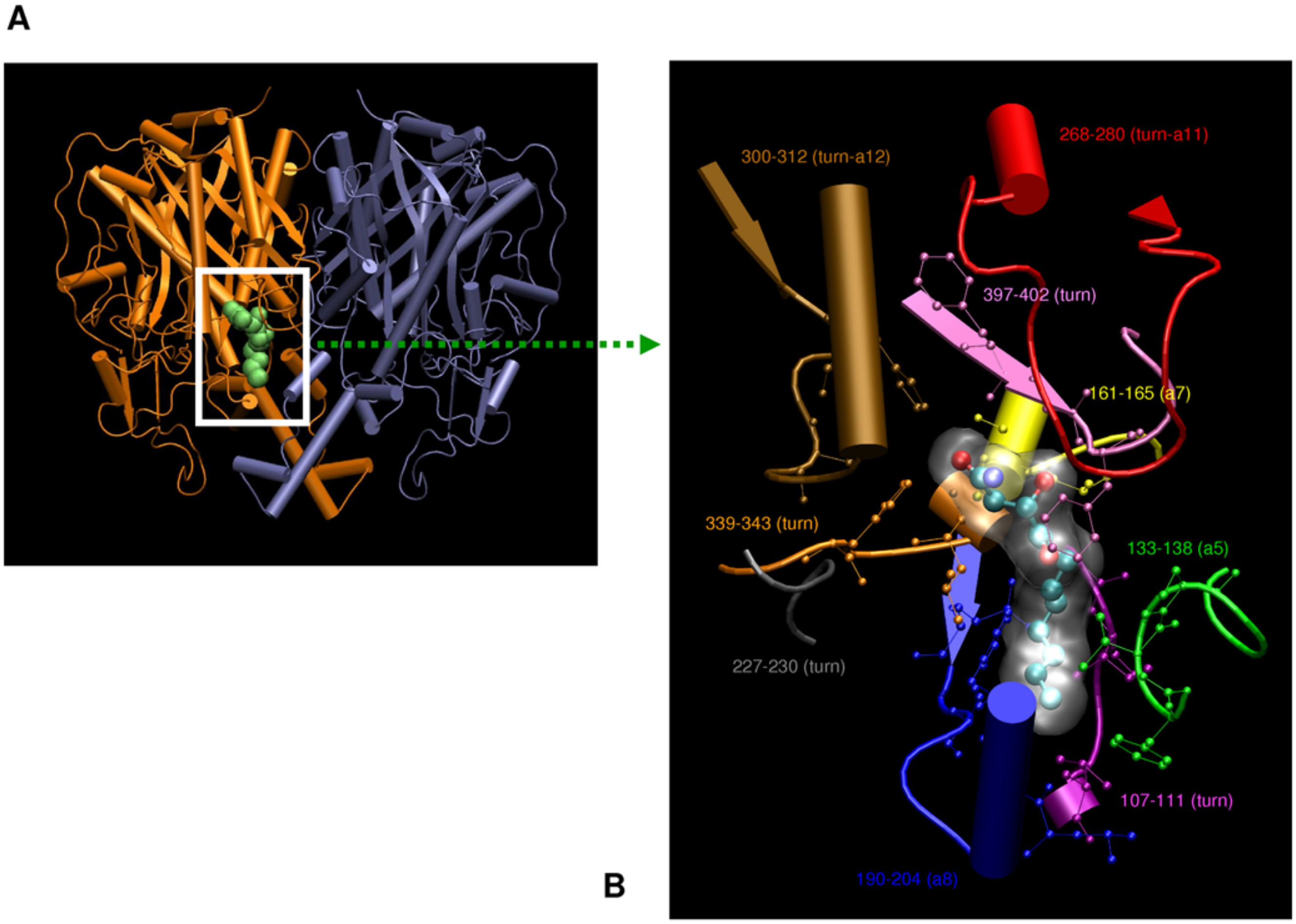


Figure 2. Structural template for modeling of iterative KS domains. (A) The *E. coli* KAS-II homo-dimer with ligand. (B) The backbones (secondary structural rendering) and side chains (ball and stick) of different stretches of amino acids that constitute the ligand binding cavity of *E.coli* KAS-II have been depicted in different colors.
doi:10.1371/journal.pcbi.1000351.g002

fumonisin KS domain by KS from lovastatin LDKS resulted in polyketides having short chain length [38]. Very recent experiments involving generation of altered fatty acid-polyketide hybrid products by rational manipulation of benastatin biosynthetic pathway [39] also suggest that number of chain elongations is dependent on the size of the PKS enzyme cavity. The *in silico* analysis of the sequence and structural features of iterative KS domains reported here provides a structural rationale for these experimentally observed variations in substrate specificities and further helps in identification of residues that can be specifically mutated to control the number of iterations in type-I PKSs. No experimental studies have as yet been reported on altering the number of iterations in type-I PKSs by site directed mutagenesis. The present *in silico* analysis gives crucial leads for such experiments.

Table 1. Comparison of the cavity volumes and hydrophobicities of various KS structural models with the number of iterations and product size.

		Product Size (No. of backbone carbons)	Number of iterations	Cavity Volume (Å^3)	Number of hydrophobic residues	Hydrophobicity	Number of CLRs
FAS	Reducing	Variable	Variable	934	18	47.7	47
MSAS	Partial	8	3	180	9	14.4	24
AVILA	Partial	8	3	291	8	25.2	20
THN	Non-reducing	10	5	819	12	24.1	32
WA-NAP	Non-reducing	14	6	895	16	38	44
T-TOXIN	Reducing	40	20	1781	16	25.7	56

CLR: Cavity Lining Residues.
doi:10.1371/journal.pcbi.1000351.t001

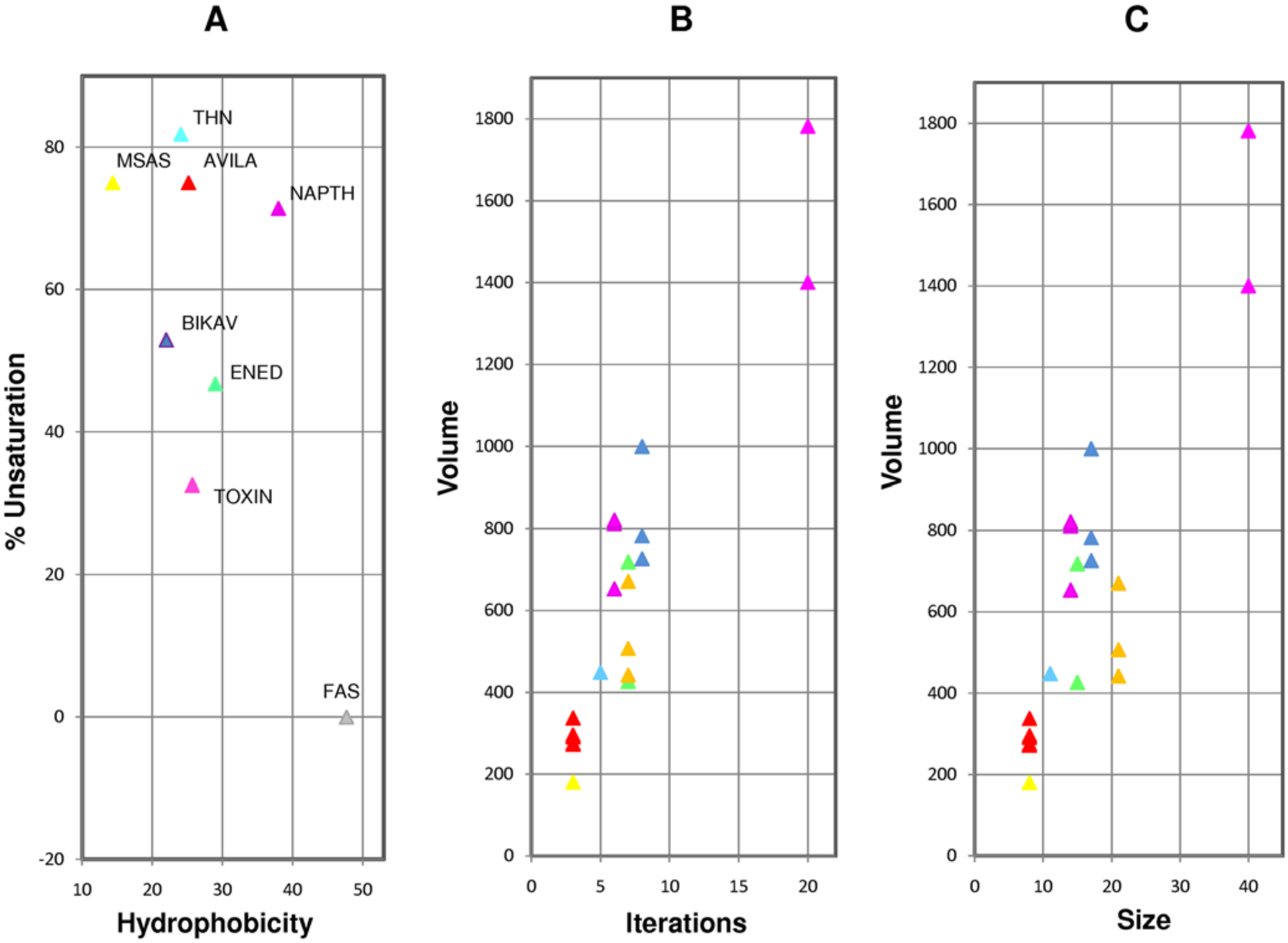
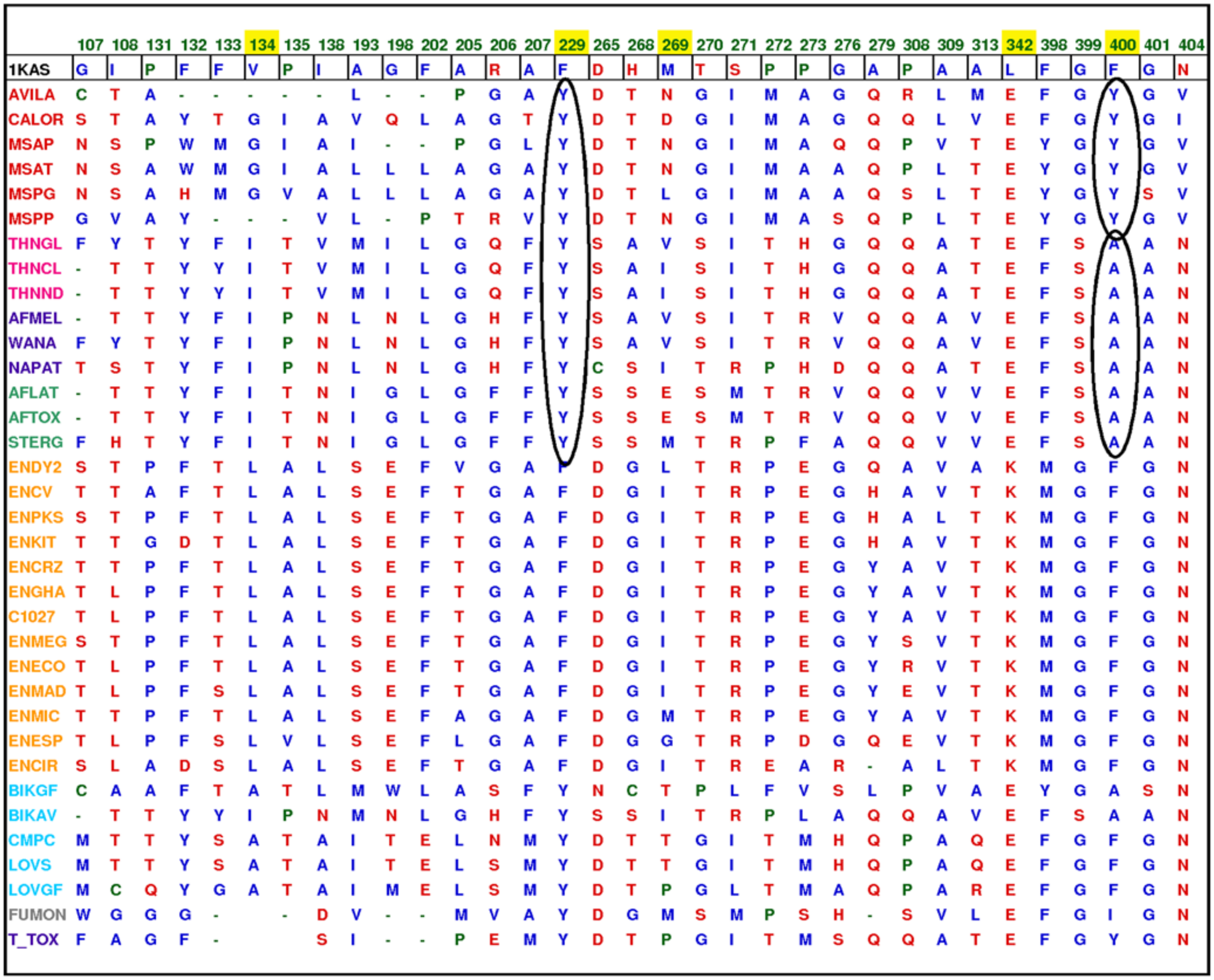


Figure 3. Variation in hydrophobicity and size of the active site cavities of various iterative KS domains. The KS domains carrying out different number of iterations have been depicted in separate colors. Points corresponding to different homology models of the same KS domain have a common color. Hydrophobicity of CLRs correlates negatively with the extent of unsaturation in the final product (A). Cavity volumes (Å^3) correlate positively with the number of iterations (B). Cavity volumes (Å^3) of iterative KS domain pockets show a positive correlation with final product size (number of backbone carbon atoms in the polyketide) (C).
doi:10.1371/journal.pcbi.1000351.g003

Predicting the order of substrate channeling in modular PKS clusters

In modular PKS clusters, the chemical structure of the product is governed by the order in which substrates are channeled between various ORFs. It has often been observed that the order of PKS ORFs during biosynthesis of a polyketide is not the same as the order of the corresponding ORFs in the genome. This complexity of module succession has been depicted in Figure S3 using schematic representation of a type I modular PKS cluster. This biosynthetic cluster has four polyketide synthase ORFs and their order in the genome is Orf1, Orf2, Orf3 and Orf4. But during the biosynthesis, Orf4 is the first to function and the product of Orf4 is transferred to Orf1. Orf2 functions at a later stage and its product is condensed with the rest of the polyketide. This inconsistency between ordering of ORFs in the genome and the order of substrate channeling is a commonly observed phenomenon, as is evident from the simocyclinone [40], nanchangmycin [41], microcystin [42], pimaricin, rapamycin and nystatin biosynthetic clusters. The prediction of the correct order of substrate channeling is essential for *in silico* identification of polyketide products of uncharacterized modular PKS clusters. Therefore, deciphering the cognate combination of ORFs in a modular PKS cluster from the large number of theoretically possible non-cognate combinations has been the major bottleneck in formulating predictive rules for *in silico* identification of polyketide products. Hence, we attempted to investigate whether predictive rules based on specificity of interaction between ORFs can be formulated for deciphering the correct order of substrate channeling in an uncharacterized PKS cluster.

Several experimental studies have suggested that inter protein interactions in modular PKSs are mediated by specific recognition between docking domains or the so called 'interpolypeptide linker' regions [24,25,29]. The amino acid stretches N-terminus to the first KS domain and C-terminus to the last ACP domain are referred as inter polypeptide linkers or docking domains. These have been extensively studied and it has been proposed that, the C-terminal (Cter) docking domains specifically pair with the N-terminal (Nter) docking domains of the succeeding ORF to facilitate cross-talk between the consecutive ORFs. Structural elucidation [29] of the cognate docking domains from erythromycin PKS (DEBS) has revealed that, unlike conventional linker sequences which join protein domains covalently within polypeptides, these docking domain regions are not non-structured, but adopt a relatively compact four helix bundle structure. It has been



	107	108	131	132	133	134	135	138	193	198	202	205	206	207	229	265	268	269	270	271	272	273	276	279	308	309	313	342	398	399	400	401	404
1KAS	G	I	P	F	F	V	P	I	A	G	F	A	R	A	F	D	H	M	T	S	P	P	G	A	P	A	A	L	F	G	F	G	N
AVILA	C	T	A	-	-	-	-	-	L	-	-	P	G	A	Y	D	T	N	G	I	M	A	G	Q	R	L	M	E	F	G	Y	G	V
CALOR	S	T	A	Y	T	G	I	A	V	Q	L	A	G	T	Y	D	T	D	G	I	M	A	G	Q	Q	L	V	E	F	G	Y	G	I
MSAP	N	S	P	W	M	G	I	A	I	-	-	P	G	L	Y	D	T	N	G	I	M	A	Q	Q	P	V	T	E	Y	G	Y	G	V
MSAT	N	S	A	W	M	G	I	A	L	L	L	A	G	A	Y	D	T	N	G	I	M	A	A	Q	P	L	T	E	Y	G	Y	G	V
MSPG	N	S	A	H	M	G	V	A	L	L	L	A	G	A	Y	D	T	L	G	I	M	A	A	Q	S	L	T	E	Y	G	Y	S	V
MSPP	G	V	A	Y	-	-	-	V	L	-	P	T	R	V	Y	D	T	N	G	I	M	A	S	Q	P	L	T	E	Y	G	Y	G	V
THNGL	F	Y	T	Y	F	I	T	V	M	I	L	G	Q	F	Y	S	A	V	S	I	T	H	G	Q	Q	A	T	E	F	S	A	A	N
THNCL	-	T	T	Y	Y	I	T	V	M	I	L	G	Q	F	Y	S	A	I	S	I	T	H	G	Q	Q	A	T	E	F	S	A	A	N
THNND	-	T	T	Y	Y	I	T	V	M	I	L	G	Q	F	Y	S	A	I	S	I	T	H	G	Q	Q	A	T	E	F	S	A	A	N
AFMEL	-	T	T	Y	F	I	P	N	L	N	L	G	H	F	Y	S	A	V	S	I	T	R	V	Q	Q	A	V	E	F	S	A	A	N
WANA	F	Y	T	Y	F	I	P	N	L	N	L	G	H	F	Y	S	A	V	S	I	T	R	V	Q	Q	A	V	E	F	S	A	A	N
NAPAT	T	S	T	Y	F	I	P	N	L	N	L	G	H	F	Y	C	S	I	T	R	P	H	D	Q	Q	A	T	E	F	S	A	A	N
AFLAT	-	T	T	Y	F	I	T	N	I	G	L	G	F	F	Y	S	S	E	S	M	T	R	V	Q	Q	V	V	E	F	S	A	A	N
AFTOX	-	T	T	Y	F	I	T	N	I	G	L	G	F	F	Y	S	S	E	S	M	T	R	V	Q	Q	V	V	E	F	S	A	A	N
STERG	F	H	T	Y	F	I	T	N	I	G	L	G	F	F	Y	S	S	M	T	R	P	F	A	Q	Q	V	V	E	F	S	A	A	N
ENDY2	S	T	P	F	T	L	A	L	S	E	F	V	G	A	F	D	G	L	T	R	P	E	G	Q	A	V	A	K	M	G	F	G	N
ENCV	T	T	A	F	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	H	A	V	T	K	M	G	F	G	N
ENPKS	S	T	P	F	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	H	A	L	T	K	M	G	F	G	N
ENKIT	T	T	G	D	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	H	A	V	T	K	M	G	F	G	N
ENCRZ	T	T	P	F	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	Y	A	V	T	K	M	G	F	G	N
ENGHA	T	L	P	F	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	Y	A	V	T	K	M	G	F	G	N
C1027	T	L	P	F	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	Y	A	V	T	K	M	G	F	G	N
ENMEG	S	T	P	F	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	Y	S	V	T	K	M	G	F	G	N
ENECO	T	L	P	F	T	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	Y	R	V	T	K	M	G	F	G	N
ENMAD	T	L	P	F	S	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	P	E	G	Y	E	V	T	K	M	G	F	G	N
ENMIC	T	T	P	F	T	L	A	L	S	E	F	A	G	A	F	D	G	M	T	R	P	E	G	Y	A	V	T	K	M	G	F	G	N
ENESP	T	L	P	F	S	L	V	L	S	E	F	L	G	A	F	D	G	G	T	R	P	D	G	Q	E	V	T	K	M	G	F	G	N
ENCIR	S	L	A	D	S	L	A	L	S	E	F	T	G	A	F	D	G	I	T	R	E	A	R	-	A	L	T	K	M	G	F	G	N
BIKGF	C	A	A	F	T	A	T	L	M	W	L	A	S	F	Y	N	C	T	P	L	F	V	S	L	P	V	A	E	Y	G	A	S	N
BIKAV	-	T	T	Y	Y	I	P	N	M	N	L	G	H	F	Y	S	S	I	T	R	P	L	A	Q	Q	A	V	E	F	S	A	A	N
CMPC	M	T	T	Y	S	A	T	A	I	T	E	L	N	M	Y	D	T	T	G	I	T	M	H	Q	P	A	Q	E	F	G	F	G	N
LOVS	M	T	T	Y	S	A	T	A	I	T	E	L	S	M	Y	D	T	T	G	I	T	M	H	Q	P	A	Q	E	F	G	F	G	N
LOVGF	M	C	Q	Y	G	A	T	A	I	M	E	L	S	M	Y	D	T	P	G	L	T	M	A	Q	P	A	R	E	F	G	F	G	N
FUMON	W	G	G	G	-		-	D	V	-	-	M	V	A	Y	D	G	M	S	M	P	S	H	-	S	V	L	E	F	G	I	G	N
T_TOX	F	A	G	F	-			S	I	-	-	P	E	M	Y	D	T	P	G	I	T	M	S	Q	Q	A	T	E	F	G	Y	G	N

Figure 4. List of residues lining the active site pockets of KS domains in various iterative PKS clusters. For clarity, positions that have completely invariant residues (for e.g. the catalytic triad) or positions with a high number of gaps have been removed from this table. The highlighted positions have been discussed in detail in the text, and are likely to govern the carbon chain length in different iterative PKSs. The two crucial positions, 229 and 400 have been circled.

doi:10.1371/journal.pcbi.1000351.g004

proposed that, this four helix bundle structure is the core fold of cross-talk [29] between ORFs of modular PKS clusters. These structures have been termed inter protein 'docking domains' to emphasize that they are responsible for the recognition and subsequent docking between successive protein modules. The C-terminal docking domain is reported to contain three helices (hereafter named helix 1, 2 and 3) whereas the N-terminal docking domain contains a single longer helix (hereafter named helix 4). This docking domain complex is a symmetrical dimer, consisting of two independent structural units called domain A and domain B. Domain A is an unusual intertwined α-helical bundle comprising helices 1 and 2. Domain B is also an α-helical bundle but with an entirely different topology and it comprises helix 3 (from Cter) and helix 4 (from Nter). Thus the actual docking interaction occurs in domain B, via several pairs of charged residues and a conserved set of hydrophobic residues. However, it has been proposed that, out of these various interacting residues, two pairs of appropriately placed charged residues at critical positions on the docking interface, form a kind of 'docking code' for DEBS [29] (Figure S4). When DEBS1 docks against DEBS2, the charges at these positions give rise to favorable interactions. However, in case of non-cognate combinations between DEBS1 and DEBS3, the resulting charge interactions are repulsive. The availability of DEBS docking domain structure provided us the opportunity to test, whether such a code exists in other PKS systems as well. We have carried out a structure based analysis of docking domain sequences to investigate if rules for identification of cognate ORF combination can be formulated based on key interactions found in DEBS docking domain structure.

It may be noted that, based on bioinformatics analysis of docking domains in type I modular PKS proteins, Broadhurst *et al* [29] had also proposed that DEBS-like docking domain structures would be present in other type I modular PKS clusters and they govern the cross-talk between ORFs. Since secondary structure analysis by Broadhurst *et al* [29] had clearly demonstrated propensity of docking domain sequences for four helix bundle structure similar to DEBS docking domain, inter polypeptide contacts were extracted for both cognate and non-cognate pairs of

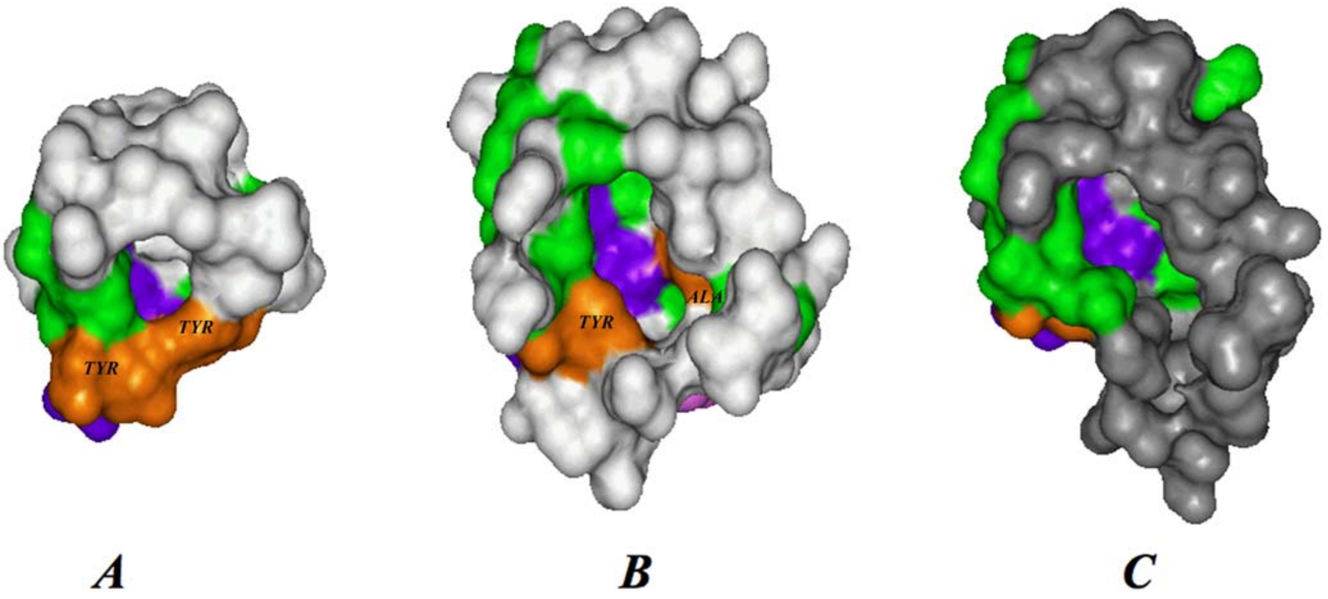


Figure 5. Functionally important cavity lining residues of two types of iterative KS domains. MSAS (A) and NAP (B). The cavities of the models have been shown in surface rendering. Each model has been superimposed with the structural template. The two orange residues correspond to the positions 229 and 400, which together block the downward flow of the MSAS cavity. One of these residues is an Ala in case of the intermediate NAP-type cavity and this allows the cavity to flow downwards. These cavities are actually buried inside the protein, and residues forming the top layer have been removed for clarity. (C) The internal topology of the structural template, *E. coli* KAS-II protein cavity has been depicted for reference. The surface has been colored such that the catalytic triad is in purple, regions which are invariant among the different iterative KS domains, are in green. Thus the differences in the cavity shapes arise from residues lying in the grey region of the depicted cavity surface. The cavity is completely buried, but the top layer of residues has been removed for clarity of the figure.
doi:10.1371/journal.pcbi.1000351.g005

ORFs in various modular PKSs using the DEBS docking domain structure as a template. Since recent studies [16,29,43] suggest that PKS docking domains fall into at least three different phylogenetic classes, our assumption regarding docking domains from various phylogenetic groups adopting similar structural folds requires further justifications. It is well known that for a given protein family, structure is conserved to a much larger extent than sequence [44,45]. There are many examples of proteins adopting similar three dimensional structural fold even in absence of detectable sequence similarity [44,45]. Recently available structures [46] of mammalian type I FAS proteins also show remarkably high similarity to structures PKS protein domains even if they share only a limited sequence homology. Therefore, our assumption regarding myxobacterial PKS 'docking domains' adopting structural folds similar to docking domains from actinomycetes is not unreasonable. Hence, we extracted crucial interacting residues for various docking domain pairs based on alignment with DEBS docking domain structure. Figure 6 shows the alignment of cognate pairs of various PKS docking domain sequences with DEBS docking domain structure. The interacting residue pairs obtained from this alignment were ranked as favorable, unfavorable or neutral as per a simple scoring scheme (Table S1). The interactions between a pair of oppositely charged amino acids or between a pair of hydrophobic amino acids were ranked as favourable, while electrostatic repulsions between a pair of charged amino acids was called unfavourable. On the other hand, interactions between any other amino acid pairs, specifically the interactions between charged and hydrophobic amino acids was ranked as neutral. It may be noted that, this simplistic scoring scheme has been defined based on types of amino acid contacts found in interfaces of protein-protein complexes [47]. A total of 66 cognate pairs of docking domain sequences were checked for the two pairs of positions which give rise to favorable electrostatic interactions in the docking domain structure. Out of these, 54 pairs of ORFs were found to have at least one residue pair with favorable interaction. Moreover, there was no cognate pair where both of these interactions were unfavorable. Thus it can be concluded that cognate pairing of ORFs does generate energetically favorable contacts.

Since a good docking code interaction was observed in more than 80% cases, we investigated if these crucial inter polypeptide contact pairs could be used to predict the correct order of module succession in a given modular PKS. If all possible combinations of ORFs in a PKS cluster are considered together, there would be only one biosynthetically correct order of ORFs. This correct combination would in turn have a set of all cognate interfaces and therefore, the highest number of favorable interactions. The remaining combinations of ORFs would be incorrect and accordingly, they would have varying numbers of non-cognate interfaces, thus resulting in unfavorable interactions. It may be added here that, the identity of the first and last ORFs can usually be established by the presence of an initiating loading module and the terminal TE domain respectively. The presence of a very short C-terminal sequence beyond the conserved TE domain can also be used as a criterion for identification of the last module. Figure 7 shows the example of the Spinosad biosynthetic cluster which has ten modules arranged in five ORFs. These five ORFs can be combined in six different ways if the first and last ORFs are fixed. Each of the six combinations would have four interfaces. All the interfaces were scanned for favorable, unfavorable or neutral interactions at the positions corresponding to the DEBS docking code. As can be seen in Figure 7, the correct order of ORFs has the highest number of favorable interactions and no repulsive interaction at any of its interfaces. In contrast, each of the remaining five combinations has at least two repulsive interactions, and thus can be rejected in comparison with the correct combination.

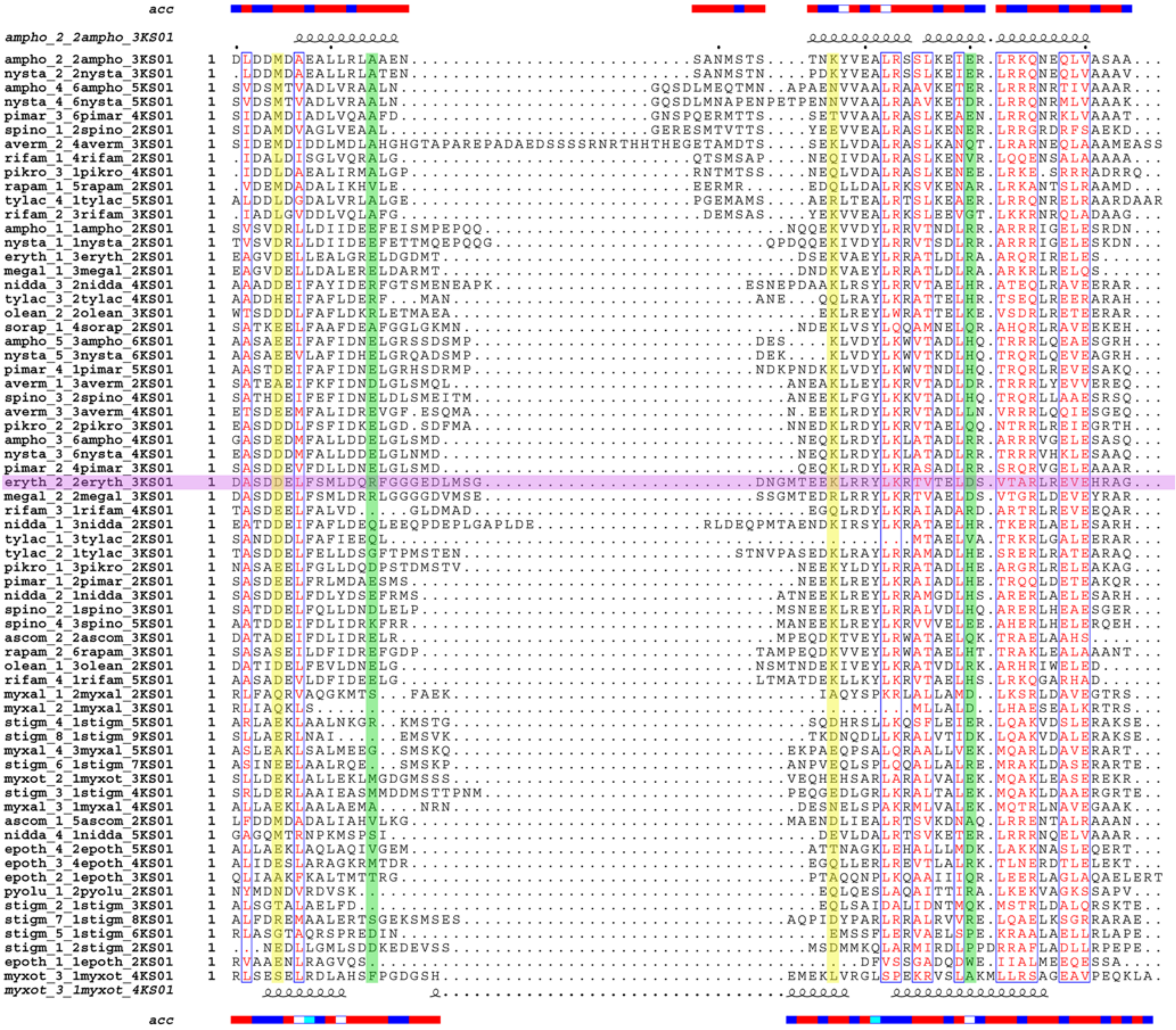
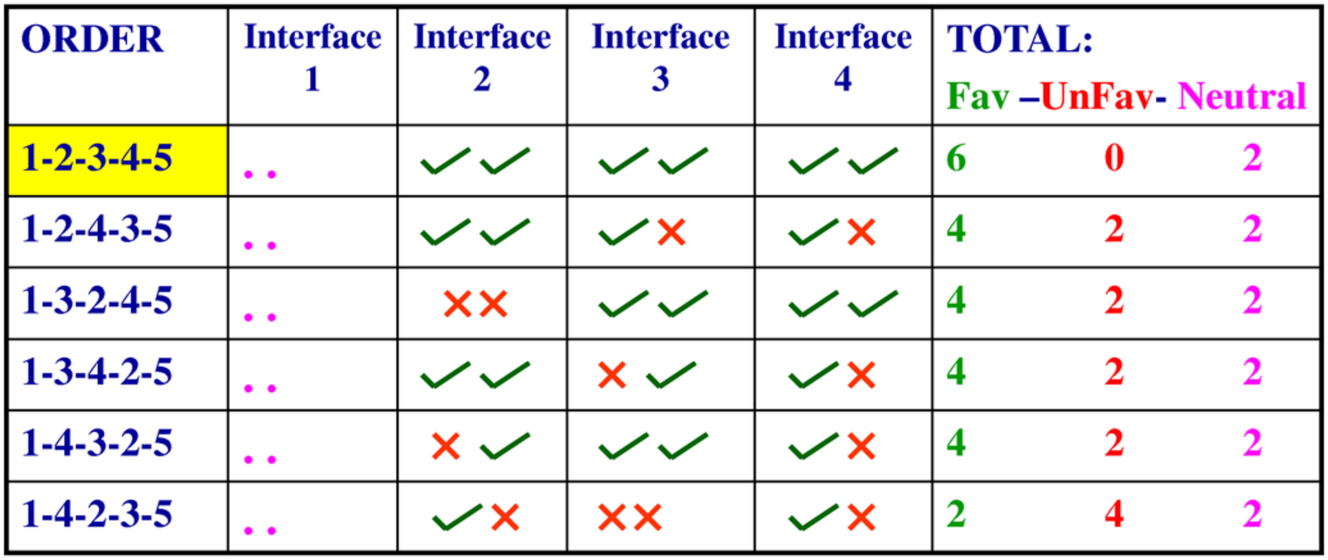


Figure 6. A structure based sequence alignment of the docking domains from various PKS clusters. Helix 3 and helix 4 were concatenated before secondary structure prediction. ESPript service [89] from the predict protein server was used for structural based sequence alignment of docking domains. The N-terminus docking domain consists of the sequence stretch extending from N-termini to the beginning of the first KS domain, while the C-terminus docking domain extends from the end of the last ACP domain to the C-terminus of the PKS protein. Inter polypeptide contacts were extracted using the DEBS NMR structure as a template. The two pairs of interacting residues which constitute the docking code have been highlighted in green and yellow respectively. The reference sequence of DEBS docking domains is highlighted in purple color.
doi:10.1371/journal.pcbi.1000351.g006

A total of 39 characterized PKS clusters were analyzed in this manner to test the validity of this assumption. For a representative set of PKS clusters, Figure 8 shows in tabular format, the number of favorable, unfavorable and neutral contacts in the cognate combination and also the number of non-cognate combinations having a score better, equal or worse compared to the cognate combination. As can be seen from Figure 8, in several modular PKS clusters unfavourable interactions are present. However, the number of unfavourable interactions is much smaller than the favourable or neutral interactions present in the cognate interfaces. Thus analysis of cognate inter polypeptide contacts in 17 modular PKS clusters suggest that, both the interactions need not be favourable for effective docking domain interactions. However, non-cognate interfaces have more number of unfavourable interactions. Hence, there are relatively few non-cognate combinations having a score better than cognate combination. In ten out of 17 PKS clusters, no non-cognate combination has better score than the cognate combination. Even though there are non-cognate combinations having scores equal to cognate combination, the cognate combination can still be ranked among top few in these 10 cases. In case of four other PKS clusters, there are a significant number of non-cognate combinations having score higher then the cognate combination. However, the cognate combination can still be ranked within top 20% of all possible combinations. For example, in case of nanchangmycin 480 non-cognate possibilities have better score than cognate, 239 have scores equal to the cognate combination. Thus the cognate combination is ranked in top 720 combinations. However, the total number of combina-



ORDER	Interface 1	Interface 2	Interface 3	Interface 4	TOTAL: Fav	–UnFav-	Neutral
1-2-3-4-5	● ●	✓✓	✓✓	✓✓	6	0	2
1-2-4-3-5	● ●	✓✓	✓✗	✓✗	4	2	2
1-3-2-4-5	● ●	✗✗	✓✓	✓✓	4	2	2
1-3-4-2-5	● ●	✓✓	✗ ✓	✓✗	4	2	2
1-4-3-2-5	● ●	✗ ✓	✓✓	✓✗	4	2	2
1-4-2-3-5	● ●	✓✗	✗✗	✓✗	2	4	2

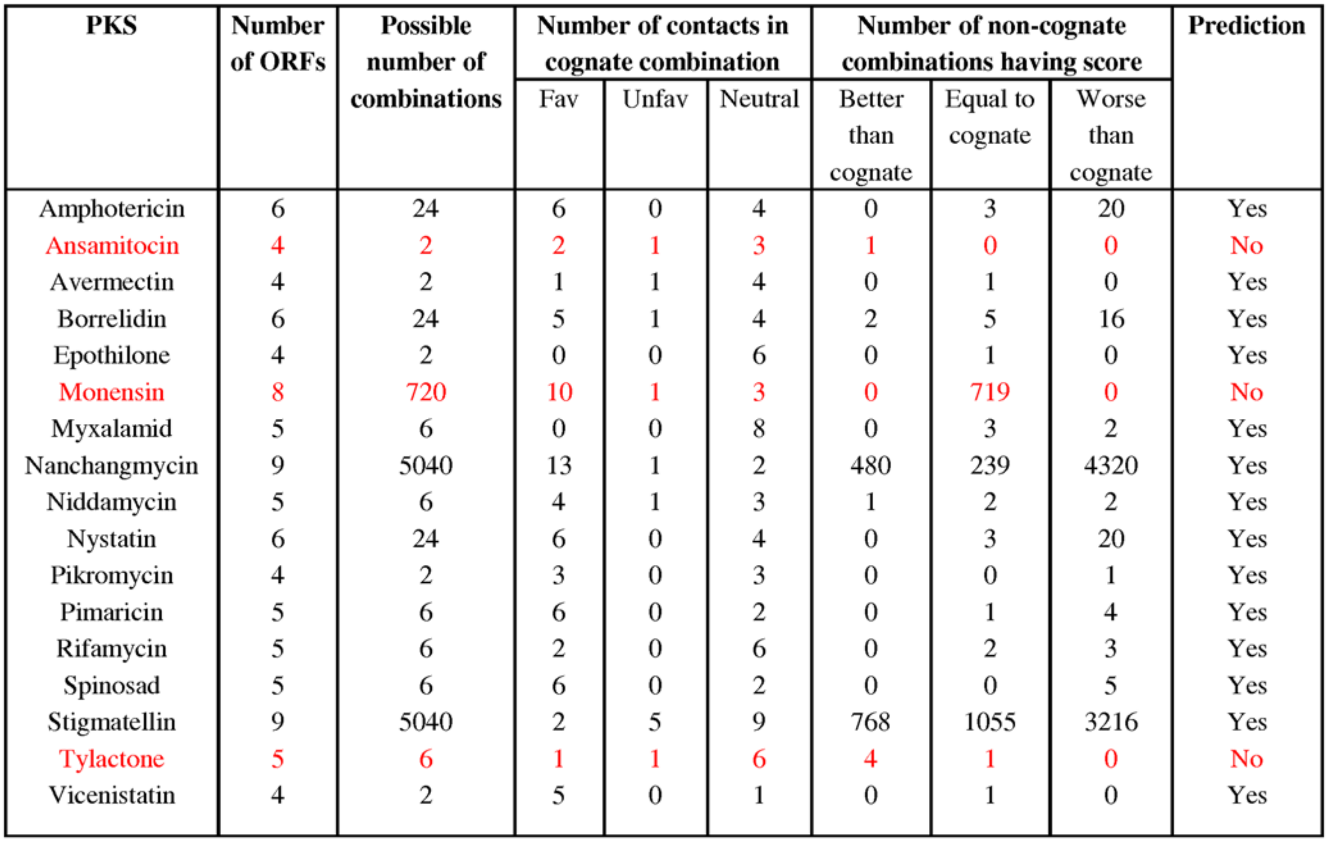
Figure 7. List of various combinatorial possibilities for the order of substrate channeling in the Spinosad modular PKS cluster. The Spinosad PKS has five ORFs which can be arranged in six different combinations, if the identity of the first and last ORF is fixed. This has been shown in the first column, where the native or correct order of ORFs has been highlighted. Each combination has four possible interfaces and each interface has been scored for two pairs of critical contacts. These two interactions can be favorable (green tick mark) or unfavorable (red cross mark) or neutral (pink dot). The last column shows the total number and type of contacts. The combination of ORFs with the highest number of favorable contacts and lowest number of unfavorable contacts is assigned as the best scorer. As can be seen, the native combination is the highest scorer in this case.
doi:10.1371/journal.pcbi.1000351.g007

torial possibilities is 5040. Therefore, our computational method ranks the cognate combination in top 14% in case of nanchangmycin PKS cluster. It is important to note that, despite the large number of combinatorial possibilities, prediction based on docking domain sequences alone is able to reject a sufficiently high number of non-cognate combinations. Thus, our results on analysis of docking domain sequences indicate that, in more than 80% of the cases the cognate order of substrate channeling can be predicted correctly. However, we must clarify that, 'correct prediction' would mean eliminating significant number of non-cognate combinations and restricting the cognate combination to a relatively smaller number of possibilities. Such a relaxed definition of 'correct prediction' can be justified by the fact that, we are using a simple prediction method involving few crucial contacting residues rather than all the interactions present in the docking domain structure. Secondly, we are not taking into account role of other catalytic domains in preventing chain elongation in case of non-cognate associations.

Even though very recent theoretical studies [5,16] have attempted to predict physical interaction between PKS proteins based on analysis of co-evolution of docking domain sequences, the prediction accuracy for order of substrate channeling has either not been studied in detail [16] or found to be low in cases involving clusters consisting of more than four ORFs [5]. However, in contrast to these purely sequence based methods, we have used a structure based approach. Using the conserved core structure of the docking domain as template, we have extracted crucial interacting residues which were suggested earlier by Broadhurst *et al* [29] to be determinants of specificity of inter subunit interactions. Exploitation of this crucial information in our study probably helps in improvement of prediction accuracy. Identification of specific interacting residue pairs also make the predictions easily amenable to experimental testing by site directed mutagenesis approach. Recent experimental studies [30,31] have further established the feasibility of altering specificity of inter subunit interactions based on manipulation of putative interacting residues in the docking domain frame work. Apart from helping in deciphering the chemical structure of final polyketide product, our computational analysis of "docking code" in cognate and non-cognate interacting pairs in experimentally characterized modular PKS cluster can also provide knowledge base for fruitfully combining non-cognate ORF pairs for generation of novel aglycone structures. Our analysis of such interacting residues in docking domains of a mycobacterial PKS protein involved in biosynthesis of mycoketide has led to the discovery of a completely novel "Modularly iterative" mechanism of polyketide biosynthesis [48]. However, we must clarify that, apart from interactions between N-terminal and C-terminal docking domains of PKS proteins, the substrate specificity of various catalytic domains would also have a role in preventing chain elongation in case of non-cognate associations of PKS ORFs. Similarly, interactions between ACP and downstream KS will also discriminate non-cognate associations. In this work, we have only addressed the role of docking domains.

Discussion

We have demonstrated that, the KS domains can be successfully classified into various functional subfamilies with high prediction accuracy using their HMM profiles. Structural modeling of the active site pockets of various iterative KS domains has revealed that certain key residues in the active site pocket can potentially control the size of final product by governing the total number of iterations. This result is in agreement with recent experiments [38,39] which report cavity volume being a major determinant of substrate specificity of fungal PKSs. The major highlight of our work is that programmed iteration by fungal polyketide synthases may be rationally controlled by site directed mutagenesis of certain specific residues. These results also demonstrate that the number of chain extension reactions catalyzed by an iterative PKS protein



PKS	Number of ORFs	Possible number of combinations	Number of contacts in cognate combination			Number of non-cognate combinations having score			Prediction
			Fav	Unfav	Neutral	Better than cognate	Equal to cognate	Worse than cognate	
Amphotericin	6	24	6	0	4	0	3	20	Yes
Ansamitocin	4	2	2	1	3	1	0	0	No
Avermectin	4	2	1	1	4	0	1	0	Yes
Borrelidin	6	24	5	1	4	2	5	16	Yes
Epothilone	4	2	0	0	6	0	1	0	Yes
Monensin	8	720	10	1	3	0	719	0	No
Myxalamid	5	6	0	0	8	0	3	2	Yes
Nanchangmycin	9	5040	13	1	2	480	239	4320	Yes
Niddamycin	5	6	4	1	3	1	2	2	Yes
Nystatin	6	24	6	0	4	0	3	20	Yes
Pikromycin	4	2	3	0	3	0	0	1	Yes
Pimaricin	5	6	6	0	2	0	1	4	Yes
Rifamycin	5	6	2	0	6	0	2	3	Yes
Spinosad	5	6	6	0	2	0	0	5	Yes
Stigmatellin	9	5040	2	5	9	768	1055	3216	Yes
Tylactone	5	6	1	1	6	4	1	0	No
Vicenistatin	4	2	5	0	1	0	1	0	Yes

Figure 8. Result of the docking code analysis. The first two columns depict a PKS cluster and its corresponding number of ORFs. The third column shows the total number of ORF combinations possible, of which only one is the correct (or native) order. All possible combinations were tested for the presence of two critical interactions. The fourth and fifth columns have been further divided into three sub-columns each. The fourth column shows the interaction score (favorable, unfavorable and neutral) for the correct order of ORFs. The fifth column depicts the number of non-native combinations which resulted in a score that was better than, same or worse than native. Rows colored red depict the cases where this prediction method failed.

doi:10.1371/journal.pcbi.1000351.g008

can be predicted by computing the cavity volume of the active site pocket of its KS domain. This represents a major advance towards prediction of the polyketide products of iterative PKS proteins.

We have analyzed the docking domain sequences of various modular PKS clusters in detail to investigate if information contained in the docking domain sequences can be used to identify the correct order for channeling of substrates. Using the recently available NMR solution structure [29] of the docking domains from the erythromycin biosynthetic cluster as template, inter polypeptide contacts were analyzed for various types of cognate and non-cognate pairing of ORFs in various modular PKS clusters. Our investigation revealed that, cognate pairing of ORFs always generated energetically favorable inter polypeptide contacts, while in majority of cases non-cognate pairing resulted in energetically unfavorable contacts. The results of our benchmarking on known modular PKS clusters indicated that, using such inter polypeptide contact analysis, it is possible to narrow down the number of possible choices for the cognate order of substrate channeling. Thus our analysis of docking domain sequences would help in predicting the final polyketide products of modular PKS clusters.

In summary, the current work demonstrates that, *in silico* analysis of experimentally characterized PKS clusters can not only enhance our understanding of mechanistic polyketide biosynthesis, it helps in formulating rules for predicting, whether a given PKS protein is modular or iterative, the order of substrate channeling for modular PKSs, and the number of chain extension reactions catalyzed by iterative PKSs. Hence, our results can aid in identifying metabolic products of uncharacterized PKS clusters found in newly sequenced genomes.

Methods

KS dataset

In addition to the PKS gene clusters cataloged in the NRPS-PKS server, additional modular PKS clusters that were used for this analysis are ansamitocin [49], albicidin [50], *Bacillus subtilis* PKS, coronafacic acid, compactin CDKS [51], lovastatin LDKS [52], geldanamycin [53], leinamycin [54], lankacidin [55], microcytin (from two organisms) [56,57], monensin [58], nanchangmycin [41], pederin [59], mupirocin [60], ta1 [61], bleomycin [62] and yersiniabactin [63]. The experimentally characterized fungal type I iterative PKS clusters used in this analysis are aflatoxin [64], avilamycin [65], bikaverin [35], C-1027 [66], calicheamicin (has two type I PKSs) [67], compactin [51], lovastatin [52], fumonisin [68], MSAS from four organisms [69–71], sterigmatocystin [72], THN from five organisms [73–76], T-toxin [77] and napthopyrone [78]. To this data, we added sequences analyzed in a previous phylogenetic analysis of fungal [79] type-I PKSs.

KS subfamilies

Profile HMM analysis [33] was carried out by HMMER package. The available KS dataset was divided into five different subfamilies. Apart from the major clusters of iterative and modular KS domains, the KS domain phylogenetic dendrogram showed further clustering into subfamilies like enediynes and non-enediynes within the iterative cluster. Similarly, modular KS domains have three clusters corresponding to pure modular PKSs, hybrid NRPS-PKSs and trans-AT systems. The enediyne family of antibiotics is structurally characterized by the enediyne core, a unit consisting of two acetylenic groups conjugated to a double bond or incipient double bond within the nine-membered or ten-membered ring. The enediyne cores bear no structural resemblance to any characterized polyketides, but precursor labeling experiments have unambiguously established that they are derived minimally from eight head-to-tail acetate units [80]. Natural products of hybrid peptide-polyketide origin have been known for a long time. These are metabolites that are assembled from amino acid and carboxylic acid precursors by hybrid NRPS-PKS gene clusters in which an NRPS-bound growing peptidyl intermediate is further elongated by a PKS module or vice versa [81]. Trans-AT clusters are also referred to as the AT-less clusters. These are complex PKSs where a single AT protein functions in trans- and charges the ACP domains of all the modules in the cluster [20]. Since the modular PKSs often have several KS domains on the same ORF, for building Hidden Markov Models of various subfamilies repartitioning of the various data sets into training and test set was done based on individual ORFs, rather than polyketide clusters or KS domains.

Modeling of iterative KS domains and analysis of their active site pockets

The various iterative KS domains were modeled using comparative modeling approach. The structural templates were identified by BLAST search against PDB or by using threading approach. Threading analyses were done using a local version of Threader package [82] (downloaded from the PSIPRED protein prediction server site) to identify the structural templates for modeling various KS domains. The various KS domains have been modeled using fatty acid KAS structure as template, which show only about 20% sequence identity with polyketide KS domains. However, availability of several structures of thiolase fold indicates that even at this low sequence identity, two KS proteins can adopt very similar structures. Since the overall active site architecture is conserved in this class of enzymes, our structural predictions are likely to be reliable even at low sequence identity between target and template. The crystal structure of the *act* KS-CLF protein and recently reported structure of DEBS KS have revealed that modular as well as iterative polyketide KS domains also adopt a thiolase fold, thus validating our assumptions.

Models of various polyketide KS domains were built using a local version of modeller V6.2 [83]. Structural mapping, ligand construction and pocket architecture visualization were done using different modules of InsightII package. The active site pockets of iterative KS domains were compared in terms of their hydrophobicity and cavity volumes to understand how binding pocket residues control chemical structure of the polyketide product. Cavity volumes were calculated using CASTp [84]. Only those cavities which contained the catalytic triad residues were chosen from the CASTp output for comparison across various models of a given KS domain. The cavity lining residues (CLRs) were identified from the selected CASTp pockets. The total number and total hydrophobicity of hydrophobic CLRs was tabulated for comparison with the FAS structural template. Hydrophobicity was calculated using Kyte and Doolittle's protein hydropathy scale [85]. Since cavity identification is often sensitive to small changes in orientation of residues, all the above mentioned parameters were calculated from at least five different homology models for the same sequence. Structural alignment of various KS structures was done using Combinatorial Extension (CE) server [86]. Visualization was also done using VMD [87].

Analysis of docking domains

Secondary structure propensities of various docking domain sequences were derived from the PredictProtein server [88]. ESPript service [89] from the predict protein server was used for structure based sequence alignment of docking domains. Interacting residues for each docking domain pair was identified by aligning their sequences with the docking domain structure. For each interface, the interacting residue pairs obtained from this alignment were ranked as favorable, unfavorable or neutral as per a simple scoring scheme (Table S1). A given combinatorial arrangement of a set of ORFs in a PKS cluster was assigned a score based on the favorable, unfavorable or neutral contacts present in all the interfaces. All the combinatorial possibilities were scored for each modular PKS cluster and score of the cognate combination was compared with scores of various non-cognate arrangements. The computational tool for carrying out inter subunit contact analysis involving docking domains and predicting the order of substrate channeling in modular PKS clusters is available as web server at http://www.nii.res.in/pred_pks_orf_order.html.

Supporting Information

Figure S1 Dendrogram of active site residues from all KS domains
Found at: doi:10.1371/journal.pcbi.1000351.s001 (0.13 MB DOC)

Figure S2 Superposition of backbones of iterative KS domain models on structural templates
Found at: doi:10.1371/journal.pcbi.1000351.s002 (0.24 MB DOC)

Figure S3 Genomic order vs biosynthetic order
Found at: doi:10.1371/journal.pcbi.1000351.s003 (0.09 MB DOC)

Figure S4 The four helix bundle structure of DEBS docking domain
Found at: doi:10.1371/journal.pcbi.1000351.s004 (0.21 MB DOC)

Table S1 Scoring scheme for docking domain interactions
Found at: doi:10.1371/journal.pcbi.1000351.s005 (0.07 MB DOC)

Acknowledgments

Authors thank Director, NII for encouragement and support.

Author Contributions

Conceived and designed the experiments: GY RSG DM. Performed the experiments: GY. Analyzed the data: GY RSG DM. Wrote the paper: GY RSG DM.

References

1. Liou GF, Khosla C (2003) Building-block selectivity of polyketide synthases. Curr Opin Chem Biol 7: 279–284.
2. Yadav G, Gokhale RS, Mohanty D (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J Mol Biol 328: 335–363.
3. Yadav G, Gokhale RS, Mohanty D (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domains. Nucleic Acids Res 31: 3654–3658.
4. Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res 32: W405–W413.
5. Minowa Y, Araki M, Kanehisa M (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J Mol Biol 368: 1500–1517.
6. Challis GL, Ravel J, Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chem Biol 7: 211–224.
7. Stachelhaus T, Mootz HD, Marahiel MA (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetetases. Chem Biol 6: 493–505.
8. Trivedi OA, Arora P, Vats A, Ansari MZ, Tickoo R, et al. (2005) Dissecting the mechanism and assembly of a complex virulence mycobacterial lipid. Mol Cell 17: 631–643.
9. Simeone R, Constant P, Guilhot C, Daffe M, 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.
10. Lautru S, Deeth RJ, Bailey LM, Challis GL (2005) Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat Chem Biol 1: 265–269.
11. Wilkinson B, Micklefield J (2007) Mining and engineering natural-product biosynthetic pathways. Nat Chem Biol 3: 379–386.
12. Bergmann S, Schumann J, Scherlach K, Lange C, Brakhage AA, et al. (2007) Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. Nat Chem Biol 3: 213–217.
13. Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, et al. (2007) The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem Biol 14: 53–63.
14. Van Lanen SG, Shen B (2006) Microbial genomics for the improvement of natural product discovery. Curr Opin Microbiol 9: 252–260.
15. Nguyen T, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, et al. (2008) Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nat Biotechnol 26: 225–233.
16. Thattai M, Burak Y, Shraiman BI (2007) The origins of specificity in polyketide synthase protein interactions. PLoS Comput Biol 3: e186. doi:10.1371/journal.pcbi.0030186.
17. Zazopoulos E, Huang K, Staffa A, Liu W, Bachmann BO, et al. (2003) A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat Biotechnol 21: 187–190.
18. McAlpine JB, Bachmann BO, Piraee M, Tremblay S, Alarco AM, et al. (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.
19. Khosla C, Gokhale RS, Jacobsen JR, Cane DE (1999) Tolerance and specificity of polyketide synthases. Annu Rev Biochem 68: 219–253.
20. Shen B (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Curr Opin Chem Biol 7: 285–295.
21. Walsh CT (2008) The chemical versatility of natural-product assembly lines. Acc Chem Res 41: 4–10.
22. Wenzel SC, 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.
23. Wenzel SC, Muller R (2007) Myxobacterial natural product assembly lines: fascinating examples of curious biochemistry. Nat Prod Rep 24: 1211–1224.
24. Kumar P, Li Q, Cane DE, Khosla C (2003) Intermodular communication in modular polyketide synthases: structural and mutational analysis of linker mediated protein-protein recognition. J Am Chem Soc 125: 4097–4102.
25. Gokhale RS, Tsuji SY, Cane DE, Khosla C (1999) Dissecting and exploiting intermodular communication in polyketide synthases. Science 284: 482–485.
26. Wu N, Cane DE, Khosla C (2002) Quantitative analysis of the relative contributions of donor acyl carrier proteins, acceptor ketosynthases, and linker regions to intermodular transfer of intermediates in hybrid polyketide synthases. Biochemistry 41: 5056–5066.
27. Tsuji SY, Cane DE, Khosla C (2001) Selective protein-protein interactions direct channeling of intermediates between polyketide synthase modules. Biochemistry 40: 2326–2331.
28. Wu N, Tsuji SY, Cane DE, Khosla C (2001) Assessing the balance between protein-protein interactions and enzyme-substrate interactions in the channeling of intermediates between polyketide synthase modules. J Am Chem Soc 123: 6465–6474.
29. Broadhurst RW, Nietlispach D, Wheatcroft MP, Leadlay PF, Weissman KJ (2003) The structure of docking domains in modular polyketide synthases. Chem Biol 10: 723–731.
30. Weissman KJ (2006) The structural basis for docking in modular polyketide biosynthesis. ChemBioChem 7: 485–494.
31. Weissman KJ (2006) Single amino acid substitutions alter the efficiency of docking in modular polyketide biosynthesis. ChemBioChem 7: 1334–1342.
32. Weissman KJ, Muller R (2008) Protein-protein interactions in multienzyme megasynthetases. ChemBioChem 9: 826–848.
33. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
34. Graziani S, Vasnier C, Daboussi MJ (2004) Novel polyketide synthase from Nectria haematococca. Appl Environ Microbiol 70: 2984–2988.
35. Linnemannstons P, Schulte J, del Mar Prado M, Proctor RH, Avalos J, et al. (2002) The polyketide synthase gene pks4 from Gibberella fujikuroi encodes a key enzyme in the biosynthesis of the red pigment bikaverin. Fungal Genet Biol 37: 134–148.
36. Moffitt MC, Neilan BA (2003) Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. J Mol Evol 56: 446–457.
37. Keatinge-Clay AT, Maltby DA, Medzihradszky KF, Khosla C, Stroud RM (2004) An antibiotic factory caught in action. Nat Struct Mol Biol 11: 888–893.
38. Zhu X, Yu F, Li XC, Du L (2007) Production of dihydroisocoumarins in Fusarium verticillioides by swapping ketosynthase domain of the fungal iterative polyketide synthase Fum1p with that of lovastatin diketide synthase. J Am Chem Soc 129: 36–37.
39. Xu Z, Schenk A, Hertweck C (2007) Molecular analysis of the benastatin biosynthetic pathway and genetic engineering of altered fatty acid-polyketide hybrids. J Am Chem Soc 129: 6022–6030.
40. Trefzer A, Pelzer S, Schimana J, Stockert S, Bihlmaier C, et al. (2002) Biosynthetic gene cluster of simocyclinone, a natural multihybrid antibiotic. Antimicrob Agents Chemother 46: 1174–1182.
41. Sun Y, Zhou X, Liu J, Bao K, Zhang G, et al. (2002) 'Streptomyces nanchangensis', a producer of the insecticidal polyether antibiotic nanchangmycin and the antiparasitic macrolide meilingmycin, contains multiple polyketide gene clusters. Microbiology 148: 361–371.
42. Rouhiainen L, Vakkilainen T, Siemer BL, Buikema W, Haselkorn R, et al. (2004) Genes coding for hepatotoxic heptapeptides (microcystins) in the cyanobacterium Anabaena strain 90. Appl Environ Microbiol 70: 686–692.
43. Richter CD, Nietlispach D, Broadhurst RW, Weissman KJ (2008) Multienzyme docking in hybrid megasynthetases. Nat Chem Biol 4: 75–81.
44. Thornton JM, Orengo CA, Todd AE, Pearl FM (1999) Protein folds, functions and evolution. J Mol Biol 293: 333–342.
45. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006) Structural diversity of domain superfamilies in the CATH database. J Mol Biol 360: 725–741.
46. Maier T, Leibundgut M, Ban N (2008) The crystal structure of a mammalian fatty acid synthase. Science 321: 1315–1322.
47. Halperin I, Wolfson H, Nussinov R (2004) Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. Structure 12: 1027–1038.
48. Chopra T, Banerjee S, Gupta S, Yadav G, Anand S, et al. (2008) Novel intermolecular iterative mechanism for biosynthesis of mycoketide catalyzed by a bimodular polyketide synthase. PLoS Biol 6: e163. doi:10.1371/journal.pbio.0060163.
49. Yu TW, Bai L, Clade D, Hoffmann D, Toelzer S, et al. (2002) The biosynthetic gene cluster of the maytansinoid antitumor agent ansamitocin from Actinosynnema pretiosum. Proc Natl Acad Sci U S A 99: 7968–7973.
50. Huang G, Zhang L, Birch RG (2001) A multifunctional polyketide-peptide synthetase essential for albicidin biosynthesis in Xanthomonas albilineans. Microbiology 147: 631–642.
51. Abe Y, Suzuki T, Ono C, Iwamoto K, Hosobuchi M, et al. (2002) Molecular cloning and characterization of an ML-236B (compactin) biosynthetic gene cluster in Penicillium citrinum. Mol Genet Genomics 267: 636–646.
52. Hendrickson L, Davis CR, Roach C, Nguyen DK, Aldrich T, et al. (1999) Lovastatin biosynthesis in Aspergillus terreus: characterization of blocked mutants, enzyme activities and a multifunctional polyketide synthase gene. Chem Biol 6: 429–439.
53. Rascher A, Hu Z, Viswanathan N, Schirmer A, Reid R, et al. (2003) Cloning and characterization of a gene cluster for geldanamycin production in Streptomyces hygroscopicus NRRL 3602. FEMS Microbiol Lett 218: 223–230.
54. Cheng YQ, Tang GL, Shen B (2002) Identification and localization of the gene cluster encoding biosynthesis of the antitumor macrolactam leinamycin in Streptomyces atroolivaceus S-140. J Bacteriol 184: 7013–7024.
55. Mochizuki S, Hiratsu K, Suwa M, Ishii T, Sugino F, et al. (2003) The large linear plasmid pSLA2-L of Streptomyces rochei has an unusually condensed gene organization for secondary metabolism. Mol Microbiol 48: 1501–1510.
56. Tillett D, Dittmann E, Erhard M, von Dohren H, Borner T, et al. (2000) Structural organization of microcystin biosynthesis in Microcystis aeruginosa PCC7806: an integrated peptide-polyketide synthetase system. Chem Biol 7: 753–764.

57. Tanabe Y, Kaya K, Watanabe MM (2004) Evidence for recombination in the microcystin synthetase (mcy) genes of toxic cyanobacteria Microcystis spp. J Mol Evol 58: 633–641.
58. Oliynyk M, Stark CB, Bhatt A, Jones MA, Hughes-Thomas ZA, et al. (2003) Analysis of the biosynthetic gene cluster for the polyether antibiotic monensin in Streptomyces cinnamonensis and evidence for the role of monB and monC genes in oxidative cyclization. Mol Microbiol 49: 1179–1190.
59. Piel J (2002) A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. Proc Natl Acad Sci U S A 99: 14002–14007.
60. El-Sayed AK, Hothersall J, Cooper SM, Stephens E, Simpson TJ, et al. (2003) Characterization of the mupirocin biosynthesis gene cluster from Pseudomonas fluorescens NCIMB 10586. Chem Biol 10: 419–430.
61. Paitan Y, Alon G, Orr E, Ron EZ, Rosenberg E (1999) The first gene in the biosynthesis of the polyketide antibiotic TA of Myxococcus xanthus codes for a unique PKS module coupled to a peptide synthetase. J Mol Biol 286: 465–474.
62. Shen B, Du L, Sanchez C, Edwards DJ, Chen M, et al. (2001) The biosynthetic gene cluster for the anticancer drug bleomycin from Streptomyces verticillus ATCC15003 as a model for hybrid peptide-polyketide natural product biosynthesis. J Ind Microbiol Biotechnol 27: 378–385.
63. Miller DA, Luo L, Hillson N, Keating TA, Walsh CT (2002) Yersiniabactin synthetase: a four-protein assembly line producing the nonribosomal peptide/polyketide hybrid siderophore of Yersinia pestis. Chem Biol 9: 333–344.
64. Feng GH, Leonard TJ (1995) Characterization of the polyketide synthase gene (pksL1) required for aflatoxin biosynthesis in Aspergillus parasiticus. J Bacteriol 177: 6246–6254.
65. Weitnauer G, Muhlenweg A, Trefzer A, Hoffmeister D, Sussmuth RD, et al. (2001) Biosynthesis of the orthosomycin antibiotic avilamycin A: deductions from the molecular analysis of the avi biosynthetic gene cluster of Streptomyces viridochromogenes Tu57 and production of new antibiotics. Chem Biol 8: 569–581.
66. Liu W, Christenson SD, Standage S, Shen B (2002) Biosynthesis of the enediyne antitumor antibiotic C-1027. Science 297: 1170–1173.
67. Ahlert J, Shepard E, Lomovskaya N, Zazopoulos E, Staffa A, et al. (2002) The calicheamicin gene cluster and its iterative type I enediyne PKS. Science 297: 1173–1176.
68. Proctor RH, Desjardins AE, Plattner RD, Hohn TM (1999) A polyketide synthase gene required for biosynthesis of fumonisin mycotoxins in Gibberella fujikuroi mating population A. Fungal Genet Biol 27: 100–112.
69. Feng GH, Leonard TJ (1998) Culture conditions control expression of the genes for aflatoxin and sterigmatocystin biosynthesis in Aspergillus parasiticus and A. nidulans. Appl Environ Microbiol 64: 2275–2277.
70. Fujii I, Ono Y, Tada H, Gomi K, Ebizuka Y, et al. (1996) Cloning of the polyketide synthase gene atX from Aspergillus terreus and its identification as the 6-methylsalicylic acid synthase gene by heterologous expression. Mol Gen Genet 253: 1–10.
71. Beck J, Ripka S, Siegner A, Schiltz E, Schweizer E (1990) The multifunctional 6-methylsalicylic acid synthase gene of Penicillium patulum. Its gene structure relative to that of other polyketide synthases. Eur J Biochem 192: 487–498.
72. Yu JH, Leonard TJ (1995) Sterigmatocystin biosynthesis in Aspergillus nidulans requires a novel type I polyketide synthase. J Bacteriol 177: 4792–4800.
73. Takano Y, Kubo Y, Shimizu K, Mise K, Okuno T, et al. (1995) Structural analysis of PKS1, a polyketide synthase gene involved in melanin biosynthesis in Colletotrichum lagenarium. Mol Gen Genet 249: 162–167.
74. Fulton TR, Ibrahim N, Losada MC, Grzegorski D, Tkacz JS (1999) A melanin polyketide synthase (PKS) gene from Nodulisporium sp. that shows homology to the pks1 gene of Colletotrichum lagenarium. Mol Gen Genet 262: 714–720.
75. Zhang A, Lu P, Dahl-Roshak AM, Paress PS, Kennedy S, et al. (2003) Efficient disruption of a polyketide synthase gene (pks1) required for melanin synthesis through Agrobacterium-mediated transformation of Glarea lozoyensis. Mol Genet Genomics 268: 645–655.
76. Feng B, Wang X, Hauser M, Kaufmann S, Jentsch S, et al. (2001) Molecular cloning and characterization of WdPKS1, a gene involved in dihydroxynaphthalene melanin biosynthesis and virulence in Wangiella (Exophiala) dermatitidis. Infect Immun 69: 1781–1794.
77. Yang G, Rose MS, Turgeon BG, Yoder OC (1996) A polyketide synthase is required for fungal virulence and production of the polyketide T-toxin. Plant Cell 8: 2139–2150.
78. Mayorga ME, Timberlake WE (1992) The developmentally regulated Aspergillus nidulans wA gene encodes a polypeptide homologous to polyketide and fatty acid synthases. Mol Gen Genet 235: 205–212.
79. Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc Natl Acad Sci U S A 100: 15670–15675.
80. Tokiwa Y, Miyushi-Saitoh M, Kobayashi H (1992) Biosynthesis of dynemicin A, a 3-ene-1,5-diyne antitumor antibiotic. J Am Chem Soc 114: 4107–4110.
81. Du L, Shen B (2001) Biosynthesis of hybrid peptide-polyketide natural products. Curr Opin Drug Discov Devel 4: 215–228.
82. Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287: 797–815.
83. Fiser A, Sali A (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 374: 461–491.
84. Binkowski TA, Naghibzadeh S, Liang J (2003) CASTp: Computed Atlas of Surface Topography of proteins. Nucleic Acids Res 31: 3352–3355.
85. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–132.
86. Shindyalov IN, Bourne PE (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 29: 228–229.
87. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14: 33–38, 27–38.
88. Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucleic Acids Res 32: W321–W326.
89. Gouet P, Courcelle E, Stuart DI, Metoz F (1999) ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics 15: 305–308.