Evolution of polyketide synthases in bacteria

Christian P. Ridley, Ho Young Lee, and Chaitan Khosla*

Departments of Chemistry, Chemical Engineering, and Biochemistry, Stanford University, Stanford, CA 94305

The emergence of resistant strains of human pathogens to current antibiotics, along with the demonstrated ability of polyketides as antimicrobial agents, provides strong motivation for understanding how polyketide antibiotics have evolved and diversified in nature. Insights into how bacterial polyketide synthases (PKSs) acquire new metabolic capabilities can guide future laboratory efforts in generating the next generation of polyketide antibiotics. Here, we examine phylogenetic and structural evidence to glean answers to two general questions regarding PKS evolution. How did the exceptionally diverse chemistry of present-day PKSs evolve? And what are the take-home messages for the biosynthetic engineer?

biosynthesis | metabolism | engineering

Polyketides are a large family of medicinally important natural products, which are formed through the condensation of acyl-thioester units such as malonyl-CoA and methylmalonyl-CoA to yield metabolites with diverse structures and biological activities. Broadly speaking, there are three separate types of polyketide synthases (PKSs) recognized in bacteria. Multimodular PKSs consist of one or more large multidomain polypeptides where the growing polyketide chain is sequentially passed from one active site to the next. Depending on the nature of their constituent catalytic domains, these megasynthases generate chemical variety and complexity in a stepwise fashion (reviewed in ref. 1). In contrast, iterative PKSs are comprised of a single set of catalysts that assemble a polyketide of controlled chain length through repetitive use of active sites (reviewed in ref. 2). In both cases, the nascent polyketide product is frequently acted on by further tailoring enzymes to generate the antibiotic. A third type of PKS (called type III PKSs) is fundamentally different in that the growing polyketide chain is never directly attached to a protein (reviewed in ref. 3). This article focuses on the evolution of only the first two PKS classes.

Bacteria, in particular Actinomycetes and Cyanobacteria, are prolific sources of polyketides, many of which possess antibiotic activity. Erythromycin, tetracycline, and amphotericin B are three well known examples of antimicrobial warfare agents from this group of bacteria that have been found useful for treating human diseases. Polyketides have also been discovered that play other roles in the environment other than to defeat microbial competitors. One such polyketide, mycolactone, is a pathogenesis-enabling immunosuppressant produced by the bacterium *Mycobacterium ulcerans*. This human pathogen is the causative agent of Buruli ulcer, but *M. ulcerans* mycolactone-negative mutants are avirulent. Addition of mycolactone to the mycolactone-negative mutants (chemical complementation) restores virulence (4). Other polyketides have been discovered that support a symbiotic relationship. One such compound, rhizoxin, is produced by a *Burkholderia* sp. symbiont of the fungus *Rhizopus* sp. Without the symbiont, the fungus cannot make the polyketide that functions to inhibit mitosis in rice plants and allow pathogenesis (5). The bryostatins are a family of polyketides that are found in marine bryozoans and are believed to be produced by bacterial symbionts (6). These compounds have demonstrated feeding deterrence against relevant predatory fish and are believed to protect the juvenile free-swimming bryozoans until they settle and develop structural defenses (7).

Given the diverse roles polyketides have been found to play in the environment, a natural question to address is how do polyketides evolve? Here, we take a look at the literature to date and present analysis in an attempt to answer this question. Our discussion focuses on selected bacterial natural products that are (at least in part) derived from both multimodular and iterative PKSs. We also discuss some implications of insights into natural evolutionary processes for efforts aimed at engineering new polyketide drugs.

Results and Discussion

Evolution of PKSs. Earlier phylogenetic studies have suggested that PKSs share a complex evolutionary history among themselves and with prokaryotic and eukaryotic fatty acid synthases (8, 9). Our objective was not to be exhaustive in our analysis, but rather to glean answers to two general questions regarding PKS evolution in the context of specific polyketide biosynthetic features. The two over-arching questions of interest to us are: How did PKSs evolve? And what are the take-home messages for the biosynthetic engineer?

Iterative PKSs. *Biosynthetic considerations.* The most widely studied family of iterative PKSs is one that is responsible for the biosynthesis of several polyfunctional aromatic antibiotics such as actinorhodin, tetracycline, doxorubicin, and frenolicin. Its core set of enzymes includes a heterodimeric ketosynthase (KS) and chain length factor (CLF), an acyl carrier protein (ACP), and a malonyl-CoA:ACP transacylase (MAT) usually recruited from fatty acid synthases. Together these four proteins comprise the minimal PKS necessary to generate a polyketide. The KS-CLF catalyzes chain elongation (and often initiation) through decarboxylative condensation of malonyl building blocks, the ACP delivers malonyl building blocks to the KS-CLF, and the MAT supplies malonyl groups to the PKS (10). The collective action of these proteins leads to formation of a highly reactive polyketide chain of defined length that is controlled by a deep pocket in the KS-CLF. This biosynthetic intermediate is then acted on by tailoring enzymes such as ketoreductases (KRs), aromatases (AROs), and cyclases (CYCs) to yield a natural product; some of these enzymes can interact with the minimal PKS so that polyketides of different chain lengths are not produced (11, 12). Additional modifications by other tailoring enzymes such as dimerases, P450 monooxygenases, methyltransferases, and glycosyltransferases can then take place to further elaborate the natural product that usually contributes significantly to the molecule's antibiotic activity (2). The presence or absence of these accessory enzymes therefore plays an important role in the diversity of aromatic polyketides found in nature.

The biosynthesis of some aromatic polyketides is primed by building blocks other than acetate units derived from malonyl-CoA decarboxylation. Examples include enterocin, which is primed by benzoate (13), benastatin, which is primed with hexanoate (14), the R1128 antibiotics, which are primed by a range of short-chain carboxylic acids (15), oxytetracycline, which is primed by malonamate (16), and daunorubicin, which is primed by propionate

(17, 18). This priming is catalyzed by initiation PKS modules, which vary depending on the priming unit. Commonly found in these initiation modules are homodimeric KSs (related to the KSIII enzymes that initiate fatty acid biosynthesis in bacteria) (17, 19) and acetyl-ACP thioesterases (AATEs; often referred to as AT homologs) (20). These KSs catalyze the first chain elongation cycle, whereas the AATEs prevent mispriming of the KS-CLF by an abundant acetate unit (20). Similar to post-PKS tailoring steps, the primer unit can also profoundly affect the biological activity of the polyketide. An example is frenolicin B, which is identical in structure to kalafungin with the exception of a propyl substituent instead of a methyl group. Frenolicin B has excellent antiparasitic activity, whereas kalafungin is much less effective (21).

Horizontal gene transfer. An earlier phylogenetic comparison of PCR-amplified KS fragments and 16S ribosomal DNA from 99 actinomycetes isolated from soil has been reported. The tree topologies for the two sets of sequence tags had little correlation with each other. Thirteen isolates with identical 16S rDNA sequences had diverse aromatic PKS genes that fell into six different antibiotic groups. Conversely, two strains with divergent 16S rDNA sequences had very similar KS sequences (22). Thus it appears that bacterial evolution and polyketide evolution are independent of each other, and that horizontal gene transfer is a driver of aromatic polyketide diversity in nature.

Based on the above biosynthetic considerations and the knowledge that horizontal gene transfer is present in these systems, we focused our analysis of evolutionary relationships on KS, CLF, priming KS and AATE homologs, and downstream tailoring enzymes found in bacteria. By subjecting these genes and corresponding proteins to phylogenetic analysis, we could assess the role of horizontal gene transfer as opposed to coevolution between genes and gain further insights into how diversity of bacterial aromatic polyketides is achieved.

KS and CLF evolution. The evolutionary relationship between KS and CLF sequences of KS-CLF heterodimers has been noted (23). Subsequent structural and mutagenesis studies have verified and elaborated on the hypothesis that the CLF arose from duplication of an ancient KS gene and subsequently evolved to fashion a well shielded binding pocket for a highly reactive polyketide chain that can grow only to a defined length (24, 25). Previous phylogenetic analysis has revealed that all KS sequences formed a distinct clade separate from the CLF sequences, suggesting that the heterodimeric pair had only evolved once. It was also noted that, within both the KS and CLF clades, the spore pigment proteins diverged early on from corresponding antibiotic PKS sequences (23). CLF sequences had a faster divergence rate than KS sequences, despite the fact that these two genes are usually adjacent to each other in gene clusters and are frequently translationally coupled. As chain length is a driver of polyketide diversity (26), this faster divergence is consistent with its pivotal role in the evolution of new antibiotic activities.

We wanted to reinvestigate the evolution of KS and CLF by incorporating a much larger dataset than was previously available (23). For our analysis, only KS and CLF sequences were included whose biosynthetic product (or chain length in the case of spore pigments) has been characterized. This product led to the inclusion of 33 KS and CLF sequences, including the putative KS and CLF sequences encoding for a four-carbon building block of the alkaloid aurachin found in a Gram-negative bacterium, *Stigmatella aurantiaca* (27). Two sets of phylogenetic analysis were performed (28). The trees depicted in Fig. 1 are Bayesian phylograms generated from DNA sequences that were aligned based on the respective protein sequences, as trees prepared with this method were judged to be most accurate but not perfect. Less accurate were parsimony trees based on protein sequences, but it was argued that if both methods gave a similar result one could be confident that the key conclusion was reliable. Therefore, we conducted both sets of analysis so that the reliability of the results could be evaluated.
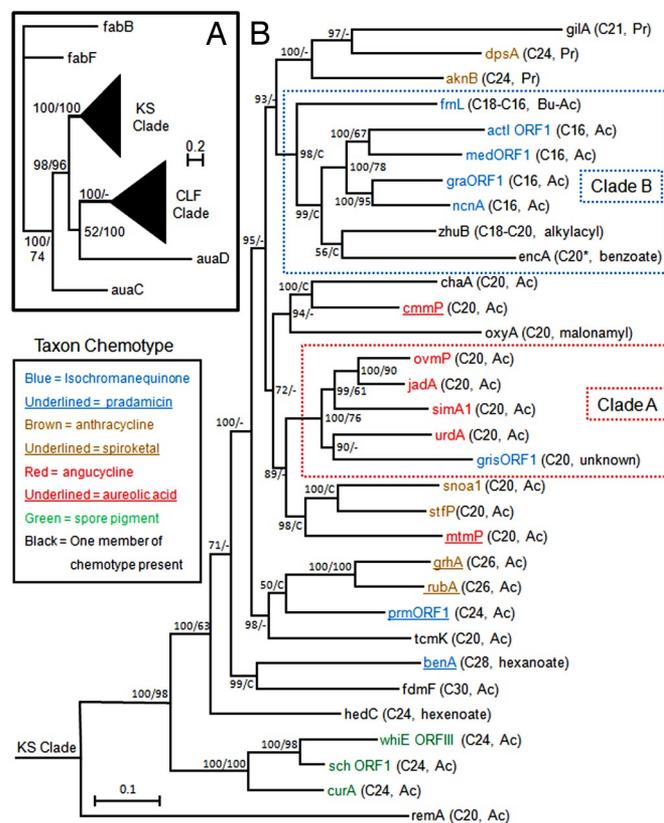

Fig. 1. Phylogenetic tree of 33 KS and CLF sequences. (*A*) The entire tree with the KS and CLF clades indicated is shown. Support for the clades is indicated by posterior probability (Bayesian)/bootstrap values (MP). (*B*) Phylogenetic tree of the KS clade found in *A*. Next to the taxon name the sizes of the respective polyketide and the primer unit are given. Ac, acetate; Pr, propionate; Bu, butyrate. Estimate of chain length is indicated by *.

The phylogram of the KS and CLF sequences is largely in agreement with the previously reported results (Fig. 1*A*). The 32 KS sequences from actinomycetes fell within a defined clade with strong support from both methods, and the aurachin CLF homolog AuaD along with the 32 other actinomycete CLF sequences fell within a supported clade. Therefore, newer data still support the premise that all KS-CLF heterodimers are descendants of a common ancestor. Interestingly, both the aurachin KS and CLF sequences fall outside of the actinomycete KS and CLF clades. The biochemical properties of these proteins need to be investigated, as they might not function as true KS-CLFs. In the KS and CLF clades [Fig. 1*B* and supporting information (SI) Fig. 6], the antibiotic resistomycin KS and CLF sequences (29) have diverged the furthest from all other KS and CLF sequences, respectively. Resistomycin is the only bacterial aromatic polyketide that has a discoid structure catalyzed by the action of several CYCs. This finding suggests that ancestral KS-CLF pairs may have served to produce antibiotics or compounds with other biological activities from which other roles such a spore pigment formation arose later. All other antibiotic gene sequences share a common ancestor with the spore pigments, as was found (23).

A question of obvious evolutionary interest in the context of aromatic polyketide biosynthesis is: At what point in their evolution did the catalytic specificity of PKSs diversify? Arguably the most important driver of polyketide structural diversity is the backbone chain length. The longer a polyketide chain, the greater are the degrees of freedom for it to undergo alternative modes of cyclization (for an example, see ref. 11). Examination of the KS and CLF trees (Fig. 1*B* and SI Fig. 6) clearly shows that the chain length
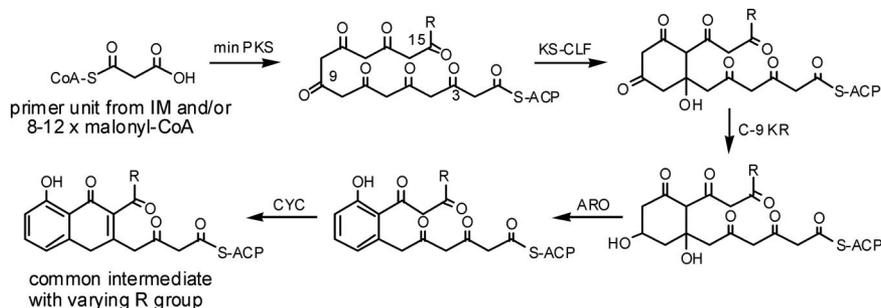
Fig. 2. Scheme of commonly found tailoring enzymes in aromatic polyketides.

specificity of minimal PKSs has diversified independently in the context of multiple antibiotic families. For example, different chain lengths can be found within separate well supported evolutionary groups such as clades A and B. If so, then it suggests that there are likely multiple routes through sequence space to engineering a PKS with desired chain-length specificity. A similar conclusion can also be drawn from the KS and CLF trees by focusing on PKSs that incorporate nonacetate primer units. Minimal PKSs that are primed with nonacetate units are found in well supported clades with acetate-primed PKSs. Thus, bimodular PKS activity is also likely to have evolved independently in multiple aromatic polyketide subfamilies and presumably contributed significantly toward optimization of antibiotic activity.

A related, but distinct, question has to do with the evolution of different chemotypes within the family of bacterial aromatic polyketides. It appears that KSs and CLFs involved in the biosynthesis of compounds of the same chemotype are usually closely related (Fig. 1B and SI Fig. 6). For example, the four angucyclines and the two spiroketals are found in well supported and distinct clades. However, some exceptions are noteworthy. For example, whereas four isochromanequinone antibiotics are located in clade B, the KS-CLF for griseusin is instead found with the angucyclines in clade A with strong support. This finding suggests that, in some instances, genetic recombination between aromatic PKS gene clusters may have given rise to antibiotics of a given chemotype with novel chain lengths. Indeed, recombining minimal PKSs and accessory biosynthetic enzymes from unrelated iterative PKS clusters has proven to be a highly productive strategy in the laboratory for the generating new aromatic polyketides (26).

Initiation module evolution. As discussed above, the minimal PKS trees suggest that initiation modules, which load nonacetate units onto the elongation KS, were likely incorporated into present-day polyketide biosynthetic pathways through horizontal gene transfer. The finding that all three KSs with propionate specificity form a defined clade in the KS tree with strong support (Fig. 1B) casts some doubt on this conclusion, as it suggests that the initiation modules and minimal PKS enzymes may be coevolving instead. We therefore wanted to investigate whether initiation KS and AATE enzymes have coevolved with the minimal PKS sequences.

Comparison of the phylogenetic trees of six elongation KS genes, which include an initiation KS (KSIII) gene in the gene cluster, revealed that coevolution of these genes is a possibility (SI Fig. 7A). Five of six KSIII sequences are found in clades similar to those present in the KS tree. A larger data set is required to definitively establish whether the KSIII genes have coevolved with the elongation KS genes, but it is an intriguing result that would not have been expected based on the data shown in Fig. 1B. Contrasting with the KSIII/KS comparison, similar analysis of eight identified AATEs with their respective KS sequences reveals no evidence for coevolution (SI Fig. 7B). The taxa in the only two clades in the AATE tree that are supported by both Bayesian and parsimony analysis have different affiliations in the KS tree. If one considers the results of these comparisons along with the data in Fig. 1, it does

seems more likely that horizontal transfer of the initiation modules between bacteria plays a larger role than coevolution. Heterologous expression of the R1128 initiation module with the actinorhodin, frenolicin, tetracenomycin, and pradimicin minimal PKSs did lead to alkyl-primed polyketides in all cases (30–32), supporting this hypothesis. Additionally, the enterocin initiation module was able to prime the actinorhodin minimal PKS with benzoate (33). Thus, horizontal transfer of initiation modules between bacteria containing different type II PKS gene clusters can be productive. It should be noted that not all initiation module and minimal PKS mix-and-match experiments are successful, as attempts to coexpress the oxytetracycline initiation module with the actinorhodin KS-CLF or the tetracenomycin KS-CLF did not lead to the production of amidated hybrid polyketides (16).

Evolution of downstream biosynthetic enzymes. Certain enzymes are frequently found associated with minimal PKSs in the biosynthetic pathways of bacterial aromatic polyketides; they can vary along with polyketide chain length and product chemotype (Fig. 2). For example, after the poyketide chain has been assembled, a specific KR reduces the C-9 carbonyl group, followed by an ARO that catalyzes a double dehydration and formation of the first aromatic ring. Another CYC then catalyzes the formation of the second ring to yield a common intermediate found in angucyclines, anthracyclines, isochromanequinones, and other chemotypes. Fourteen clusters were identified that contained these three genes, and phylogenetic analysis was performed for them in comparison with trees containing the respective KS and CLF sequences (SI Fig. 8). The gilvocarcin genes (34) were used as the outgroup for all five trees without implying that the genes are a valid evolutionary outgroup. Instead, the intent was to keep the outgroup constant for comparison purposes.

In both the KS and CLF trees, the angucycline and isochromanequinone sequences formed well supported clades, so these affiliated sequences were then used to look for possible gene transfer events that would be indicated by other genes falling inside these groups. The other sequences present in the KS and CLF trees are not located in well supported clades, leaving some doubt as to their placement, and therefore we will not discuss the incongruent clades between the various trees for these other sequences relative to each other. Two C-9 KRs were identified in the chartreusin gene cluster (35). One of these, ChaE, falls within the isochromanequinone clade. This finding suggests that this KR might have transferred into the gene cluster, perhaps from a strain making an isochromanequinone antibiotic. The other KR-, KS-, CLF-, ARO-, and CYC-encoding genes are grouped together in this cluster over a span of 6.5 kb, whereas the ChaE gene is located 11 kb away from this group. This placement suggests that this gene was incorporated from another cluster, leading to the observed duplication, and that ChaD is likely the original C-9 KR. If individual PKS genes for aromatic polyketides can be transferred between clusters, it is plausible that duplication of genes would occasionally result and these two KRs could be an example of this duplication. In the ARO tree, the medermycin ARO med ORF19 (36) is found far outside

of the clade containing the other three isochromanequinone sequences. It seems likely that this ARO has been acquired from another gene cluster. Analysis of the gene cluster shows that the ARO gene is separated by 15 kb of DNA away from the KS, CLF, C-9 KR, and CYC genes, which are located in a tight group, providing further support for this hypothesis. Because no other ARO was found in the cluster, this ARO is the only identified candidate to form the first aromatic ring in medermycin biosynthesis. Examination of the CYC tree reveals that the isochromanequinone clade is intact again, providing additional support that medermycin ARO had an outside source. These data strongly suggest that individual genes can be transferred productively from outside clusters to allow biosynthesis of different aromatic polyketides, which can lead to structural diversification or alternatively rediscovery of a bioactive chemotype in the evolution of aromatic polyketides.

Implications for the biosynthetic engineer. Overall, our analysis suggests that aromatic polyketide antibiotic diversification is dynamic, with gene transfers occurring between bacteria consisting of part or entire gene clusters. Initiation modules varying the starter unit of the aromatic polyketide can be productively exchanged, as can downstream tailoring enzymes. Chain-length specificities of the PKSs appear to evolve as is necessary for the bacteria to meet environmental pressures. Thus, by selectively incorporating initiation modules and tailoring enzymes while engineering the chain length of the minimal PKS as found in nature, biosynthetic engineers should be able to productively generate new polyketides with potentially new bioactivities more efficiently than previous efforts. Not only is this theme borne out by evolutionary history, but available data from laboratory studies also support the premise.

Multimodular PKSs. ***Biosynthetic considerations.*** The synthesis of many complex polyketide antibiotics in bacteria is catalyzed by multimodular PKSs. The core of each module consists of a KS, acyltransferase (AT), and ACP domain. Together they extend the growing polyketide chain by two carbon atoms and also generate an ACP-bound β-ketoacyl intermediate. The β-keto group can be modified by optional accessory domains, such as KR, dehydratase (DH), and enoyl reductase (ER) domains, which are typically attached to the core module. It has been suggested multimodular PKSs arose from gene and intragene module duplication, and that the prototypical PKS module shares ancestry with a vertebrate fatty acid synthase (37, 38). By understanding the nature's strategies for evolving novel multimodular PKSs, it may be possible to obtain useful clues for biosynthetic engineering in the laboratory.

Multiple mechanisms for evolution of multimodular PKSs. In addition to point mutations, three other mechanisms have been invoked for the evolution of multidomain proteins: gene and module duplication, homologous recombination, and horizontal gene transfer. Available evidence suggests that all three mechanisms have played a role in the evolution of functionally diverse multimodular PKSs.

By now there are several excellent examples in the literature that make a compelling case for the hypothesis that multimodular PKSs arose from duplication events. For example, the core KS and ACP domains of most rapamycin PKS modules have >80–85% pairwise identity, and several complete modules of the mycolactone PKS exhibit >90% pairwise identity (39). Indeed, module duplication may have been the primary mechanism underlying the evolution of all major families of polyketide antibiotics, because modules of most PKSs are more considerably similar to each other than to modules from distantly related polyketide pathways. For example, the core domains of modules from the avermectin PKS have higher sequence identity to each other than to modules from any other known synthase (40).

Sequence analysis also provides evidence for the role of homologous recombination in module diversification. For example, in case of the avermectin PKS, a previous study suggested that DH-KR domains were exchanged between modules 2 and 4 and
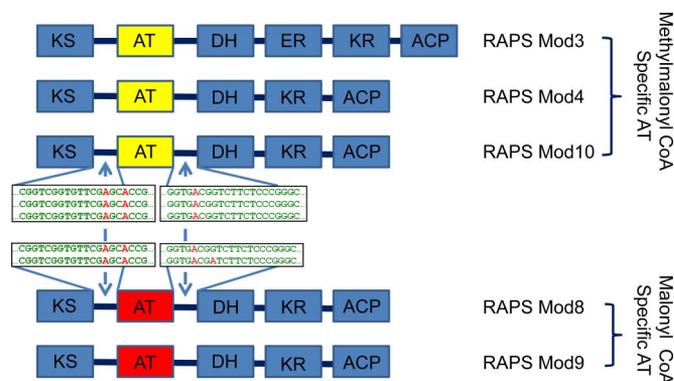

Fig. 3. Homologous recombination of AT domains in rapamycin PKS gene cluster. The sequences specified correspond to the parts of linker regions where the degree of sequence identity undergoes a marked transition.

between modules 8 and 5. Loss or gain of KR domains was also proposed in modules 3 and 4 through homologous recombination (41). It is also plausible that the AT domains of the rapamycin PKS (42) were swapped by a similar mechanism, considering the high sequence similarity of the flanking domains, especially in methylmalonyl-specific modules 3, 4, and 10 and malonyl-specific modules 8 and 9 (Fig. 3).

Last, but not least, horizontal gene transfer has also clearly played a role in PKS evolution. For example, the corresponding modules of the erythromycin and megalomicin PKSs, both of which produce 6-deoxyerythronolide B, have markedly higher sequence identity (≈75%) than with any other module from the same PKS (≈40%; SI Fig. 9). It is likely that the entire PKS was transferred en masse between these distantly related bacteria (*Saccharopolyspora erythraea* and *Micromonospora megalomicea*, respectively). A similar conclusion also emerges from the sequences of PKSs that synthesize the aglycones of closely related 16-membered macrolides such as tylosin, mycinamicin, and niddamicin. That said, it is noteworthy that the sequences of the erythromycin and megalomicin PKSs have diverged considerably (average 75% identity among pairwise modules), and the sequences of the 16-membered macrolide PKSs have diverged even further (average 50% identity among pairwise modules). This finding suggests that horizontal gene transfers between these organisms were relatively ancient events.

Evolution of AT domains. The divergence of AT domain specificity was one of the most important factors contributing to the evolution of polyketide diversity. It has been noted that malonyl-specific AT domains cluster into a distinct clade from methylmalonyl-specific AT domains (8, 40, 43), suggesting that all domains with the same specificity share a common ancestor. Taken together with the above discussion, one might conclude that nature has successfully performed AT domain substitutions on numerous occasions. For example, 12 of 14 KS and ACP domains of the rapamycin PKS share high sequence identity (70–97%), but the malonyl-specific AT domains in these modules are quite different from methylmalonyl-specific AT domains (35% identity). Because efficient engineering of AT domain substitutions in the laboratory remains an elusive challenge, it behooves us to carefully examine nature's strategy for AT domain swapping.

To further investigate the divergence of AT substrate specificity in multimodular PKSs, both Bayesian analysis of the aligned DNA sequences and maximum parsimony (MP) analysis of the aligned protein sequences were performed. Our analysis included representative malonyl-, methylmalonyl-, ethylmalonyl-, and methoxymalonyl-specific AT domains from the rapamycin, FK520, tylosin, geldanamycin, herbimycin, avermectin, niddamicin, concanamycin A, tautomycetin, and epothilone PKSs and stand-alone malonyl-specific AT proteins from the disorazol, pederin, and leinamycin
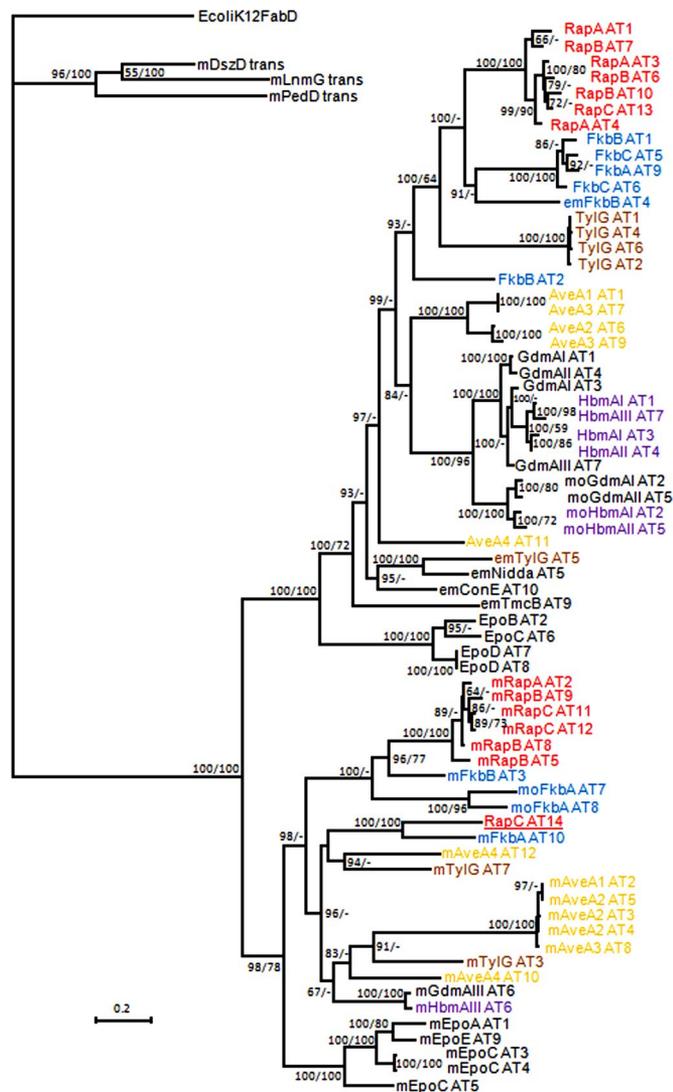
Fig. 4. Phylogenetic tree of AT gene or domain sequences. m in front of the gene indicates a malonyl CoA-specific AT, em indicates a ethylmalonyl CoA-specific AT, and mo indicates a methoxymalonyl-specific AT. The absence of a prefix implies that the AT has specificity for methylmalonyl-CoA. Support for the clades is indicated by posterior probability (Bayesian)/bootstrap values (MP).

Fig. 5. Homologous recombination hot spots for AT swaps depicted in DEBS module 5 with the KS-AT common motifs (with high sequence identity) shared by Rap modules 3, 4, 8, 9, and 10 and DEBS module 5 highlighted.

PKSs (Fig. 4). As expected, the malonyl and methylmalonyl AT groups formed distinct clades, each with strong statistical support. One noteworthy exception is the malonyl-specific RapC AT14, which is found within the methylmalonyl clade. Module 14 of the rapamycin synthase is highly unusual in its organization, as it consists of only the core KS, AT, and ACP domains with a very short AT-ACP linker of only 60 residues. In a previous study (44), it was observed that sequence changes at the C-terminal end of the AT domain could result in altered substrate specificity. It is possible that, in contrast to all other known malonyl-specific AT domains, the AT domain of module 14 evolved as a result of an intramodular deletion that also led to altered AT substrate specificity. The selective advantage, if any, of incorporating a malonyl group at this position in the polyketide chain remains to be investigated.

Although the ethylmalonyl AT domains from the FK520, concanamycin, tautomycetin, niddamycin, and tylosin PKSs lie within the methylmalonyl AT clade, implying that they evolved from methylmalonyl AT domains, more careful evaluation of the phylogenetic tree (Fig. 4) suggests that these relatively unusual AT domains may have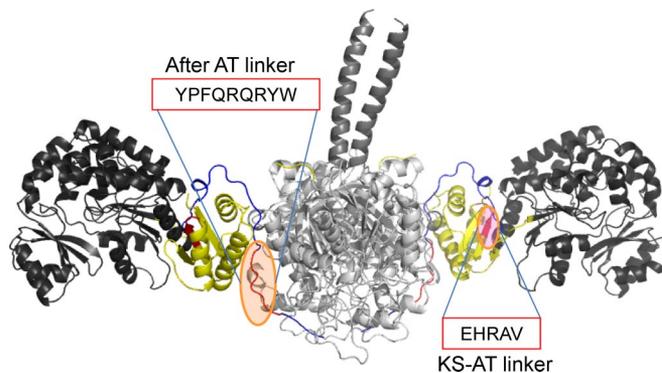 evolved on more than one occasion. For example, the TylG AT5, Nidda AT5, and ConE AT10 domains cluster into one subclade, but the FkbB AT4 and TmcB AT9 domains are more distantly related. It also appears that methoxymalonyl-specific ATs have had several progenitors, as two distinct groups of methoxymalonyl-specific AT domains are observed, one in the methylmalonyl AT clade and the other in the malonyl AT clade. Because the biosynthesis of both building blocks [especially methoxymalonyl building blocks (45)] requires elaborate pathways, it is likely that these infrequently used AT domains can evolve with comparable ease from malonyl or methylmalonyl AT domains as long as the requisite precursor is available in the cell. As AT domains evolve, it seems that natural selection for improved activity and the availability of requisite precursors allow the specificity of ATs to change over time.

In summary, because the evolution of modules with diverse building block specificity has probably involved both point mutations and domain swapping, one could analyze the modules of present-day PKSs for clues regarding how to engineer AT domain substitutions. The rapamycin PKS is a good case study, because modules with different building block specificity have islands of limited sequence identity (30% in AT domains) flanked by highly similar protein sequences (80% in modules excluding AT domains). Remarkably, in all such cases the boundaries between these mosaic module sequences lie in the structurally characterized KS-AT linker and the post-AT linker (Fig. 5). A similar conclusion emerges from comparing the sequences of modules with alternative building block specificity from the avermectin PKS (41). Interdomain linkers with defined structures have been identified in multimodular PKSs (46). Perhaps there is a key evolutionary role for these noncatalytic regions that have limited sequence similarity but a high degree of structural conservation. In essence, linkers may have evolved to maintain structural integrity notwithstanding the perturbations that arise from homologous recombination, thereby greatly enhancing the evolutionary degrees of freedom for multimodular PKSs. This hypothesis is conceptually analogous to the junctional flexibility associated with CDR3 in Ig genes, which not only contributes to antibody diversity but also provides a structurally appropriate motif for V(D)J recombination (47).

Recently, several PKSs with modules that entirely lack an AT domain have also been identified (6, 48–50). Instead, these PKSs include a stand-alone AT protein that transacylates all of the ACP domains of the PKS with malonyl building blocks. Our phylogeny analysis suggests that these stand-alone ATs comprise a distinct, more ancient clade (Fig. 4). Unlike the canonical AT domains, which have continued to evolve rapidly and use different precursors, the stand-alone AT proteins, all being exclusively malonyl-specific, have undergone less evolution and consequent divergence. Understanding how such AT-less modules evolved promises to provide clues for biosynthetic engineering of novel PKSs.

Conclusion

This analysis of PKSs reveals a class of enzymes that can be modified or amplified to increase the structural diversity of polyketide products. In this manner, bacteria capable of producing these natural products can contend with other microbes who may gain resistance to the parent polyketide and expand into new niches such as pathogenesis or symbiosis. This modification of PKSs, in combination with horizontal gene transfer events, has led to the diversity of polyketides produced by bacteria.

Nature's strategies for evolving PKSs are not haphazard, as consistent themes are apparent from the analysis presented here and elsewhere. Iterative PKSs in particular appear to achieve a lot of diversity through horizontal gene transfer mechanisms. Initiation modules can be exchanged, leading to alternate primer units, and downstream tailoring enzymes can be acquired that allow for new polyketide structures to be produced. In addition, the chain length of polyketides is a dynamic entity that appears to change under evolutionary pressures by a method comparable to site-directed mutagenesis, as just a few amino acid substitutions in the CLF protein can lead to altered chain length. Modular PKS genes appear to be transferred horizontally between bacteria similar to iterative PKS genes, but additional drivers for evolution are present. Module duplication followed by sequence divergence and acquisition of tailoring and AT domains appears to have also played a major role in the diversification of these polyketides. By following nature's lessons, it may now be possible to effectively produce new therapeutics through biosynthetic engineering in a timely fashion.

Methods: Phylogenetic Analysis

The gene sequences in this study were from the GenBank database (51); their accession numbers or the source organisms are listed in SI Tables 1 and 2. The gene sequences were aligned based on their respective protein alignments by using Mega4 (52), and the alignments were manually fine-tuned afterward based on structural and mechanistic considerations. Guided by the results of a study that assessed the accuracy of various phylogenetic methods in reconstructing evolutionary trees (28), we analyzed each gene set by using both Bayesian analysis of the aligned DNA sequences and MP analysis of the aligned protein sequences. The Bayesian phylogenetic trees were constructed by using Mr. Bayes 3.1.2 (53) with the DNA sequences partitioned into three blocks (the three codon positions). Default priors and the general time reversible model were used with a γ distribution of rate variation across sites allowing for a proportion of invariable sites. The partitions were unlinked to allow for independent estimations of likelihood. Four Markov chain Monte Carlo chains were run for enough generations (400,000 to 1.8 million), and the temperature parameter was varied as necessary so that convergence was reached (usually standard deviation of split frequencies of <0.01 were achieved). Trees were sampled every 100 generations, and the first quarter of the total trees were discarded before the analysis. The figures of the Bayesian phylograms were prepared by using Mega4 or TreeView (54).

Support for the clades found in the phylogenetic trees is indicated by posterior probability (Bayesian)/bootstrap values (MP). If the clade is not present in the MP tree, support is indicated with a − or a C if present with a bootstrap value of <50. The scale bars in the figures indicate the number of substitutions per nucleotide position, which is a measure of the evolutionary distance of two PKSs from their common ancestor.

The MP analysis of the aligned protein sequences was performed by using the PAUP 4.0b10 program (55). Heuristic searches were performed with 10–100 replicates, and sequences were added randomly. Bootstrap analysis was conducted with either 25 (Fig. 4), 100 (Fig. 1 and SI Fig. 6) or 1,000 (SI Figs. 7 and 9) replicates. Where the number of equally parsimonious trees was less than five, clades supported poorly by the bootstrap analysis (<50%) on the parsimony trees are noted in the figures. Where the number was more than five (Fig. 4), only those clades with bootstrap support >50% are indicated.

1. Fischbach MA, Walsh CT (2006) *Chem Rev* 106:3468–3496.
2. Hertweck C, Luzhetskyy A, Rebets Y, Bechtold A (2007) *Nat Prod Rep* 24:162–190.
3. Austin MB, Noel JP (2003) *Nat Prod Rep* 20:79–110.
4. Adusumilli S, Mve-Obiang A, Sparer T, Meyers W, Hayman J, Small PLC (2005) *Cell Microbiol* 7:1295–1304.
5. Partida-Martinez LP, Hertweck C (2005) *Nature* 437:884–888.
6. Sudek S, Lopanik NB, Waggoner LE, Hildebrand M, Anderson C, Liu H, Patel A, Sherman DH, Haygood MG (2007) *J Nat Prod* 70:67–74.
7. Lopanik NB, Targett NM, Lindquist N (2006) *Mar Ecol Prog Ser* 327:183–191.
8. Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) *Mol Biol Evol* 22:2027–2039.
9. Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) *Proc Natl Acad Sci USA* 100:15670–15675.
10. Reeves CD (2003) *Crit Rev Biotechnol* 23:95–147.
11. Yu T-W, Shen Y, McDaniel R, Floss HG, Khosla C, Hopwood DA, Moore BS (1998) *J Am Chem Soc* 120:7749–7759.
12. Peric-Concha N, Borovicka B, Long PF, Hranueli D, Waterman PG, Hunter IS (2005) *J Biol Chem* 280:37455–37460.
13. Piel J, Hertweck C, Shipley PR, Hunt DM, Newman MS, Moore BS (2000) *Chem Biol* 7:943–955.
14. Xu Z, Schenk A, Hertweck C (2007) *J Am Chem Soc* 129:6022–6030.
15. Marti T, Hu Z, Pohl NL, Shah AN, Khosla C (2000) *J Biol Chem* 275:33443–33448.
16. Zhang W, Ames BD, Tsai S-C, Tang Y (2006) *Appl Environ Microbiol* 72:2573–2580.
17. Bao W, Sheldon PJ, Wendt-Pienkowski E, Hutchinson CR (1999) *J Bacteriol* 181:4690–4695.
18. Räty K, Kantola J, Hautala A, Hakala J, Ylihonko K, Mäntsälä P (2002) *Gene* 293:115–122.
19. Pan H, Tsai SC, Meadows ES, Miercke LJW, Keatinge-Clay AT, O'Connell J, Khosla C, Stroud RM (2002) *Structure (London)* 10:1559–1568.
20. Tang Y, Koppisch AT, Khosla C (2004) *Biochemistry* 43:9546–9555.
21. Omura S, Tsuzuki K, Iwai Y, Kishi M, Watanabe S, Shimizu H (1985) *J Antibiot* 38:1447–1448.
22. Metsä-Ketelä M, Halo L, Munukka E, Hakala J, Mäntsälä P, Ylihonko K (2002) *Appl Environ Microbiol* 68:4472–4479.
23. Hopwood DA (1997) *Chem Rev* 97:2465–2497.
24. Tang Y, Tsai S-C, Khosla C (2003) *J Am Chem Soc* 125:12708–12709.
25. Keatinge-Clay AT, Maltby DA, Medzihradszky KF, Khosla C, Stroud RM (2004) *Nat Struct Mol Biol* 11:888–893.
26. McDaniel R, Ebert-Khosla S, Hopwood DA, Khosla C (1995) *Nature* 375:549–554.
27. Sandmann A, Dickschat J, Jenke-Kodama H, Kunze B, Dittmann E, Müller R (2007) *Angew Chem Int Ed Eng* 46:2712–2716.
28. Hall BG (2005) *Mol Biol Evol* 22:792–802.
29. Jakobi K, Hertweck C (2004) *J Am Chem Soc* 126:2298–2299.
30. Tang Y, Lee TS, Khosla C (2004) *PLoS Biol* 2:227–238.
31. Tang Y, Lee TS, Lee HY, Khosla C (2004) *Tetrahedron* 60:7659–7671.
32. Lee TS, Khosla C, Tang Y (2005) *J Am Chem Soc* 127:12254–12262.
33. Izumikawa M, Cheng Q, Moore BS (2006) *J Am Chem Soc* 128:1428–1429.
34. Fischer C, Lipata F, Rohr J (2003) *J Am Chem Soc* 125:7818–7819.
35. Xu Z, Jakobi K, Welzel K, Hertweck C (2005) *Chem Biol* 12:579–588.
36. Ichinose K, Ozawa M, Itou K, Kunieda K, Ebizuka Y (2003) *Microbiology* 149:1633–1645.
37. Cortes J, Haydoek SF, Roberts GA, Bevitt DJ, Leadlay PF (1990) *Nature* 348:176–178.
38. Donadio S, Staver MJ, McAlpine JB, Swanson SJ, Katz L (1991) *Science* 252:675–679.
39. Stinear TP, Mve-Obiang A, Small PLC, Frigui W, Pryor MJ, Brosch R, Jenkin GA, Johnson PDR, Davies JK, Lee RE, et al. (2004) *Proc Natl Acad Sci USA* 101:1345–1349.
40. Ikeda H, Nonomiya T, Usami M, Ohta T, Omura S (1999) *Proc Natl Acad Sci USA* 96:9509–9514.
41. Jenke-Kodama H, Börner T, Dittmann E (2006) *PLoS Comput Biol* 2:1210–1218.
42. Schwecke T, Aparicio JF, Molnar I, Konig A, Khaw LE, Haydock SF, Oliynyk M, Caffrey P, Cortes J, Lester JB, et al. (1995) *Proc Natl Acad Sci USA* 92:7839–7843.
43. Wu K, Chung L, Revill WP, Katz L, Reeves CD (2000) *Gene* 251:81–90.
44. Lau J, Fu H, Cane DE, Khosla C (1999) *Biochemistry* 38:1643–1651.
45. Ligon J, Hill S, Beck J, Zirkle R, Monar I, Zawodny J, Money S, Schupp T (2002) *Gene* 285:257–267.
46. Tang YY, Kim CY, Mathews II, Cane DE, Khosla C (2006) *Proc Natl Acad Sci USA* 103:11124–11129.
47. Furukawa K, Shirai H, Azuma T, Nakamura H (2001) *J Biol Chem* 276:27622–27628.
48. Piel J (2002) *Proc Natl Acad Sci USA* 99:14002–14007.
49. Cheng YQ, Tang GL, Shen B (2003) *Proc Natl Acad Sci USA* 100:3149–3154.
50. Carvalho R, Reid R, Viswanathan N, Gramajo H, Julien B (2005) *Gene* 359:91–98.
51. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007) *Nucleic Acids Res* 35:D21–D25.
52. Tamura K, Dudley J, Nei M, Kumar S (2007) *Mol Biol Evol* 24:1596–1599.
53. Ronquist F, Huelsenbeck JP (2003) *Bioinformatics* 19:1572–1574.
54. Page RDM (1996) *Comput Appl Biosci* 12:357–358.
55. Swofford DL (2003) PAUP (Sinauer, Sunderland, MA).