**Evolutionary Implications of Bacterial Polyketide Synthases**

**Holger Jenke-Kodama,* Axel Sandmann,† Rolf Müller,‡ and Elke Dittmann*†

*Humboldt University, Institute of Biology, Chausseestrasse, Berlin, Germany; and †Pharmaceutical Biotechnology, Saarland University, Saarbrücken, Germany

Polyketide synthases (PKS) perform a stepwise biosynthesis of diverse carbon skeletons from simple activated carboxylic acid units. The products of the complex pathways possess a wide range of pharmaceutical properties, including antibiotic, antitumor, antifungal, and immunosuppressive activities. We have performed a comprehensive phylogenetic analysis of multimodular and iterative PKS of bacteria and fungi and of the distinct types of fatty acid synthases (FAS) from different groups of organisms based on the highly conserved ketoacyl synthase (KS) domains. Apart from enzymes that meet the classification standards we have included enzymes involved in the biosynthesis of mycocyclic acids, polyunsaturated fatty acids (PUFA), and glycolipids in bacteria. This study has revealed that PKS and FAS have passed through a long joint evolution process, in which modular PKS have a central position. They appear to have derived from bacterial FAS and primary iterative PKS and, in addition, share a common ancestor with animal FAS and secondary iterative PKS. Furthermore, we have carried out a phylogenomic analysis of all modular PKS that are encoded by the complete eubacterial genomes currently available in the database. The phylogenetic distribution of acyltransferase and KS domain sequences revealed that multiple gene duplications, gene losses, as well as horizontal gene transfer (HGT) have contributed to the evolution of PKS I in bacteria. The impact of these factors seems to vary considerably between the bacterial groups. Whereas in actinobacteria and cyanobacteria the majority of PKS I genes may have evolved from a common ancestor, several lines of evidence indicate that HGT has strongly contributed to the evolution of PKS I in proteobacteria. Discovery of new evolutionary links between PKS and FAS and between the different PKS pathways in bacteria may help us in understanding the selective advantage that has led to the evolution of multiple secondary metabolite biosyntheses within individual bacteria.

**Introduction**

The polyketide class of natural products shows a remarkable functional and structural diversity. Apart from being toxic for microorganisms or higher eukaryotes, some of the compounds play a role in metal transport (Crosa and Walsh 2002), others are closely linked to microbial differentiation (Black and Wolk 1994; Ohnishi et al. 1999). Polyketides are classified according to the architecture of their biosynthesis enzymes. Each of the classes of polyketide synthases (PKS) resembles one of the classes of fatty acid synthases (FAS): the type I PKS possess a multidomain architecture similar to the type I FAS of fungi and animals and type II PKS carry each catalytic site on a separate protein, characteristic of FAS II found in bacteria and plants (fig. 1). Whereas usually contain monomodular iterative PKS I, the majority of bacterial PKS I consists of multiple sets of domains, or modules, that normally correspond to the number of acyl units in the product (Staunton and Weissman 2001, fig. 1). Apart from the clearly defined PKS and FAS types an increasing number of biosynthesis pathways are described in the literature that show hitherto unknown organization forms (Moss, Martin, and Wilkinson 2004). Enzymes involved in the biosynthesis of ω-3-polynsaturated fatty acids (PUFA) in *Shewanella* are authentic bacterial iterative PKS I (Metz et al. 2001, fig. 1) as well as enzymes involved in avalamycin (Gaitatzis et al. 2001), neocarzinostatin (Liu et al. 2005), and myxochromide (Wenzel et al. 2005) biosynthesis in streptomycetes and myxobacteria. Furthermore, a number of multimodular PKS I pathways are described in the literature that comprise iteratively acting modules, e.g., the biosynthesis of aureothin (He and Hertweck 2003).

Modular PKS I are predominantly found in actinobacteria, myxobacteria, pseudomonades, and cyanobacteria (Bode and Müller 2005). A minimal module is composed of a ketoacyl synthase (KS) domain, an acyltransferase (AT) domain, and an acyl carrier protein (ACP) domain. Frequently ketoreductase (KR), dehydratase (DH), and enoyl reductase (ER) domains are also embedded in the multifunctional megasynthases (fig. 1). Genetics and biochemistry of bacterial type I polyketide biosynthesis has been well investigated for the biosynthesis of the aglycone of erythromycin in *Saccharopolyspora erythrea* (Donadio et al. 1991). These findings have subsequently led to the elucidation of many PKS I pathways, in particular those involved in the formation of promising drug leads (for review, see Staunton and Weissman 2001). In bacteria, the type I PKS pathway is frequently co-occurring with a second type of natural product pathway, the nonribosomal peptide synthetases (NRPS, Shen et al. 2001). Both types of enzymes can form hybrid biosynthesis complexes, and modules of both enzyme classes can even form hybrid synthetases (Duitman et al. 1999; Paitan et al. 1999; Silakowski et al. 1999).

As striking as the number of PKS gene clusters in some bacteria is the irregular distribution of metabolites and the corresponding genes in single strains and genera in all producing families of bacteria. This has raised the hypothesis of a horizontal gene transfer (HGT) between bacterial strains. Recent phylogenetic studies of PKS I were based on the highly conserved KS domains. Kroken et al. (2003) have found evidence that fungal KS domains cluster according to the reduced or unreduced character of the polyketide products. Furthermore, KS domains from hybrid PKS/NRPS complexes form a distinct branch in phylogenetic trees (Shen et al. 2001). Piel et al. have shown that KS domains fall into a separate group when distinct
acyltransferases (so-called trans-ATs, fig. 1) are associated with PKS I systems that lack internal ATs (Piel et al. 2004). Whereas most of the studies were based on a limited set of data, Kroken et al. (2003) have presented a systematic study on the diversity and genealogy of all available fungal PKS sequences. The authors have concluded that the discontinuous distributions of orthologous PKS among fungal species can be explained by gene duplication, divergence, and
gene loss and that HGT among fungi was not necessarily involved in the evolution process.

A systematic study on the evolution of bacterial PKS is still missing. A high number of genomes of eubacteria has been completely sequenced within the last few years (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi). Biinformatic approaches are now being developed for annotation and specific analyses of the genomes. Yadav, Gokhale, and Mohanty (2003) have developed a platform for the analysis of PKS megasynthases that includes almost all current knowledge about these types of enzymes and that can be applied to dissect the arrangement of domains within these enzymes and to assign hypothetical substrate specificities of single domains (http://www.nii.res.in/nrps-pks.html). This allows a fast analysis of all PKS encoded by the microbial genomes that are completely available including those that currently cannot be assigned to a polyketide metabolite. We have chosen the AT and the KS domains for the phylogenetic study of bacterial PKS to get a conclusive picture of the evolutionary and functional relationships between domains from the different bacterial groups.

The aim of this study is to (1) investigate the evolution of bacterial PKS I and to relate it to the complex evolutionary history of the various types of FAS and PKS in the different groups of organisms, (2) systematically screen for the presence and number of PKS genes in all sequenced bacterial genomes and to test whether the number of PKS modules can be related to the genome size, (3) reveal the phylogenetic relation between PKS sequences from the different groups of bacteria, and thereby (4) assess the impact of gene duplications, gene loss, and HGT on the distribution of bacterial PKS I. A phylogenetic analysis of a complete set of functionally related domains in all groups of organisms can lead to the discovery of new evolutionary links and can help us to relate the evolution of the enzymes with the ecology and physiology of the bacteria.

Materials and Methods
Data Retrieval and Domain Analysis

The amino acid sequences of PKS I were retrieved from the National Center for Biotechnology Information microbial genome platform (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi). A BlastP search with the expected value set to the default value of 10 was performed using the protein sequence of DEBS1 from S. erythraea as the query sequence against 138 complete eubacterial, 20 complete archaeabacterial, and 3 unfinished genomes, namely, from the cyanobacterial strains Anabaena variabilis, Crocosphaera watsonii, and Nostoc punctiforme, respectively. The latter three genomes were included into the analysis to increase the data set for cyanobacteria that are known to be a rich source of secondary metabolite gene clusters (Bode and Müller 2005). All BlastP search results were inspected by eye to exclude improper sequences from the data collection. The obtained sequences were subsequently analyzed using the SEARCHPKS program (Yadav, Gokhale, and Mohanty 2003) in order to dissect the domain organization, to assign the substrate specificities, and to extract the sequences of AT and KS domains. Regarding the substrate specificity AT domains were grouped into four categories: (1) AT domains with known substrate specificities from biochemically characterized pathways, (2) AT domains with substrate specificities predicted by the SEARCHPKS program, (3) AT domains manually assigned to a substrate by analysis of amino acid residues assumed to be involved in substrate recognition, and (4) AT domains with unclear specificity. RNA sequences of the small ribosomal subunits (SSU RNA) were retrieved from the European ribosomal RNA database (http://www.psb.ugent.be/rRNA/ssu).

The amino acid sequences of FabH and FabF homologues and annotated FAS and PKS were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi).

Alignment

A total of 142 AT domains and 137 KS domains derived from the complete genome survey were subjected to a phylogenetic analysis. We furthermore included sequences of the DEBS proteins and of PKS I involved in the synthesis of the myxobacterial secondary metabolites epothilone (Sorangium cellulosum So ce90), stigmatellin, myxalamid (both Stigmatella aurantiaca Sg a15), soraphen (S. cellulosum So ce26), and pyoluteorin (Pseudomonas fluorescens Pf-5) to increase the data set for the myxobacteria and γ-proteobacteria and also to increase the number of well-characterized protein sequences and the reliability of tree reconstruction.

Amino acid alignments were created using ClustalW (Thompson, Higgins, and Gibson 1994) and adjusted manually using the MacClade program version 4.03 (W. R. Maddison and W. P. Maddison 2000). For the adjustment procedure the secondary structure of selected domains were predicted by means of the PSIPRED server (McCuffin, Bryson, and Jones 2000), and the prediction results were compared to the crystal structure of the FabD (Serre et al. 1995) and the FabF proteins (Moche et al. 1999) from Escherichia coli for the alignment of AT and KS domains, respectively. This was done to ensure correct alignment of secondary structure elements. The FabD and the FabF proteins from several bacterial strains served as outgroups in the analysis of AT and KS domains, respectively. The alignments are provided as supplementary material (Supplementary Material online).

Phylogenetic Analyses

We used different methods to reconstruct phylogenies for the amino acid alignment. For a reconstruction based on Bayesian statistics we used the MrBayes program version 3 (Huelsenbeck 2000). The Bayesian inference method employed the JTT amino acid replacement model (Jones, Taylor, and Thornton 1992) and a gamma distribution to represent among-site rate heterogeneity (JTT + γ). A discrete gamma distribution with four categories was assumed to approximate the continuous function. In the case of KS domains and proteins taken from FAS and PKS Metropolis-coupled Markov chain Monte Carlo analysis (MCMC) was performed with 1.5 million generations and four independent chains. The Markov chain was sampled every 100 generations. In the case of AT and KS domains MCMC analysis was performed with four million generations and four independent chains. As before, the Markov chain was sampled
every 100 generations. Convergence was judged by plots of maximum likelihood (ML) scores and by using the run statistics. The MCMC analysis was assumed to have reached the convergence state if all acceptance rates for the moves in the “cold” chain were in the range 10%–70% and if the acceptance rates for the swaps between chains were also in the range 10%–70%. All trees sampled before reaching the convergence state were discarded, and the remaining trees were used to construct a consensus tree and to calculate the posterior clade probabilities.

In addition, we conducted ML, neighbor-joining (NJ), and maximum parsimony analysis. Details are given in the supplementary material (Supplementary Material online) together with the corresponding phylogenetic trees.

Estimation of the Number of Duplications, Losses, and HGT Events from Phylogenetic Trees

For the assessment of the number of putative gene duplications, gene losses, and HGT events from the phylogenetic trees we considered two different types of assumptions. Firstly, the sequence clusters in the phylogenetic tree belonging to the same bacterial group and at least partially to the same organism could have been already present in the common ancestor of the respective organisms. Alternatively, they could originate from an originally homologous sequence after speciation. In the first case, gene losses must be considered and the organism showing the highest number of gene copies determines the number of duplications. The calculated value was considered as the minimal number of duplication events explaining the distribution of sequences in the tree. In the second case, the assumption of gene losses is unnecessary. The maximum sum of duplications was therefore calculated from the sum of duplications in each organism. HGT events were deduced from anomalous distribution among bacterial groups and incongruities among the sequences in the phylogenetic trees. The direction of potential HGT was inferred by considering which bacterial groups were outnumbered in the respective clades of the tree.

Results and Discussion

Evolutionary Relationships Between PKS and FAS

Fatty acid synthesis is found ubiquitously across all groups of organisms and, thus, is likely a very ancient biochemical pathway. Because FAS and PKS use the same core of enzymatic activities (fig. 1), it is reasonable to assume an evolutionary connection between these two biosynthesis systems. To prove this hypothesis a data set was created containing a selection of KS protein sequences representing all classes of FAS and the major types of PKS found in bacteria and fungi. The Bayesian phylogenetic tree derived from these sequences reflects the long joint evolution process that FAS and PKS have passed during species development (fig. 2). Similar topologies were obtained with NJ and parsimony methods (see Supplementary Material online). Archaeabacterial KS sequences of the FabH type were chosen as the outgroup. The first two clades of the reconstructed tree comprise sequences representing the dissociative type II FAS/PKS systems found in eubacteria and plants. The two subtypes of KS involved in fatty acid biosynthesis in the majority of eubacteria, FabH and FabF (fig. 1, I), fall into distinct subclades. The latter clade splits up in a subclade including all eubacterial KS of the FabF type and a second one comprising the Kx and Kp homologous of iterative PKS II of actinobacteria (fig. 1, I and V). FabF proteins from plastids and mitochondria are located near to their eubacterial counterparts, consistent with the prokaryotic origin of these organelles. The corresponding genes were transferred to the nucleus during eukaryotic evolution. Another branch of this subclade is built up from the mycobacterial FabF homologues KasA and KasB, which are involved in the synthesis of mycolic acids, high molecular weight α-alkyl-β-hydroxy acids unique to the so-called Corynebacterium-Mycobacterium-Nocardia group (CMN group) within actinobacteria (Brennan and Nikaido 1995). Bacteria of the CMN group represent a remarkable exception within the prokaryotes because they use, like fungi and animals, a multidomain FAS I and not the type II enzymes for the de novo synthesis of their long-chain fatty acids (Kikuchi, Rainwater, and Kolattukudy 1992) (fig. 1, II). Remarkably, mycolic acids synthesis involves both enzyme systems. First, the multidomain FAS I produces medium–chain length C12 to C16 fatty acids which are then transferred to the type II system. This synthase subsequently elongates the fatty acids from the first step into the very long meromycolic acids, the precursors of mycolic acids (Schweizer and Hofmann 2004). The mycobacterial FAS I cluster in close proximity of fungal FAS I (fig. 2, II and III). This close relationship coincides with the very similar architecture of both multienzymes (fig. 1, II and III). The genome of Mycobacterium tuberculosis comprises 19 genes of probable eukaryotic origin (Gamieljen, Ptitsyn, and Hide 2002). The FAS proteins, however, were not regarded as having evolved by HGT.

The following clade contains sequences from eubacterial iterative PKS I and eubacterial glycolipid synthases (fig. 1, VI and VII). Photobacterium profundum and Shewanella oneidensis are marine bacteria capable of producing ω–3 PUFAs such as docosahexaenoic acid (22:6ω3, DHA) and eicosapentaenoic acid (20:5ω3, EPA). It was shown that in bacteria PUFAs can be synthesized by an iterative PKS I (Metz et al. 2001; Wallis, Watts, and Browse 2002). Similarly, the lipid moiety of some bacterial glycolipids is produced by iteratively acting PKS (Campbell, Cohen, and Meeks 1997). This group includes the heterocyst glycolipid synthases of nitrogen-fixing cyanobacteria. The unifying characteristic of both multienzymes is a special domain architecture comprising up to five consecutive ACP domains (KS-AT-[ACP]2–5–[KR]). The only exceptions are SgcE and NesE, iterative PKS proteins involved in the biosynthesis of the enediyne antibiotics C-1027 (Liu et al. 2002) and neocarzinostatin (Liu et al. 2005) in Streptomyces globisporus and Streptomyces carzinostaticus, respectively (fig. 1, VII).

Regarding its position in the phylogenetic tree, the class of eubacterial iterative PKS I could be the ancestor of the whole range of eubacterial modular PKS I which are combined in the next clade. One subclade contains sequences from “normal” modular PKS possessing integrated cis-AT domains in each module (fig. 1, VIII). Besides sequences from modular enzymes, this subclade includes...
the iterative orsellinic acid synthase AviM of the avilomycin pathway (Weitnauer et al. 2001) and the highly similar enzyme NcsB suggested to form the naphtoic acid moiety of neocarzinostatin (Liu et al. 2005). These two sequences form a subbranch together with the uncharacterized monomodular enzyme Pks4 from *Streptomyces avermitilis*. The iteratively acting module of the modular aureothin pathway (He and Hertweck 2003) clusters in a neighbor branch together with sequences from modular PKS of streptomycetes. The other subclade comprises modular PKS acting together with trans-AT proteins, namely, those from *Bacillus subtilis* and from the leinamycin biosynthesis cluster of *Streptomyces*.
S. atroolivaceus (Tang, Cheng, and Shen 2004) and the pederin cluster of the Paederus fuscipes symbiont (Piel 2002) (fig. 1, IX). This group was described recently as a distinct phylogenetic lineage among modular PKS (Piel et al. 2004).

The iterative PKS I from fungi form a side branch of the eubacterial modular PKS I, i.e., they are clearly more closely related to those than to the fungal FAS I (fig. I, VII). Interestingly, this group of fungal sequences includes the bacterial enzyme MchA that has recently been shown to be responsible for the formation of the aliphatic side chains of myxochromides in Streptomyces aurantiaca (Wenzel et al. 2005). The position of MchA in the tree probably indicates an HGT from fungi. The top of the tree is formed by the FAS I of animals showing a remarkable proximity to modular PKS of eubacteria (fig. 1, IV). This is an important clue that may help to solve the controversially discussed question of how the fusion type FAS found in fungi and animals evolved from the originally distinct proteins. There are fundamental biochemical differences between the fungal and animal FAS regarding the nature of the termination reactions, the cofactors used by the ER activity, and different types of AT domains (McCarthy and Hardie 1984). Additionally, the domain organization and the phylogenetic relationships suggest that independent evolutionary events may have led to the development of fungal and animal FAS systems. Whereas the mycobacterial and fungal FAS I may have evolved by protein fusion from bacterial FAS II systems, animal FAS I shares a common ancestor with PKS I. Moreover, the tree reconstruction indicates that modular PKS may be the evolutionary link between the primary iterativity apparent in the type II systems and early type I FAS and PKS and the secondary iterativity of fungal PKS I and animal FAS I that is also described for an increasing number of bacterial PKS. From the data set analyzed in this study, these conclusions can only be drawn for KS domains. Each domain in an individual PKS or FAS might possess a separate evolutionary history. To generalize the findings obtained for the KS domains it would be necessary to analyze all domain types, interdomain regions, and intron sites in the eukaryotic sequences.

Taken together, the comprehensive phylogenetic analysis of the various types of FAS and PKS from different organismic groups reveals a joint evolution process of these two important biosynthetic pathways. The PKS systems in general and the modular PKS I of bacteria in particular seem to have a central position in this evolutionary interplay. In the following sections modular PKS I of bacteria will be analyzed more deeply using a phylogenomic approach.

Distribution of PKS I Among Bacteria

We could detect PKS I genes in 27 of the 138 bacterial genomes completely sequenced at the beginning of this survey and the three unfinished genomes included in our analysis (for details see Materials and Methods) representing 21% of the total number of genomes. None of the available archaeobacterial genomes possess potential PKS sequences. Archaeabacteria lack a FabD homologue, though all other FAS II components could be detected (Pereto, Lopez-Garcia, and Moreira 2004). The corresponding AT activity in this lineage presumably has been lost early in evolution and replaced by nonhomologous enzymes. Thus, one can hypothesize that archaeabacteria could not develop PKS systems because of the missing AT necessary to “construct” them. The number of genomes containing PKS I genes varied considerably between the different eubacterial groups. Whereas none of the genomes from chlamydia and spirochaetales encoded PKS I, between 5% and 77% of the genomes assigned to the firmicutes, proteobacteria, cyanobacteria, and actinobacteria were found to possess PKS I genes. An overview about the number and distribution of PKS I in completely sequenced bacterial genomes is shown in table 1. The majority of PKS I proteins comprised a single PKS module composed of at least the KS, AT, and ACP domains. Multimodular PKS I genes were abundant in the genomes of S. avermitilis and B. subtilis str. 168. The latter strain is further exceptional among the completely sequenced bacterial genomes as it was found to encode exclusively PKS modules missing an integrated AT domain along with trans-AT proteins. A complete list of the domain arrangement of PKS I analyzed in this study is available as supplementary information (Supplementary Material online). In the majority of cases, those bacterial strains encoding PKS I were also found to encode NRPS. However, whereas in actinobacteria these two types of enzyme classes are mostly encoded on separate gene clusters, in proteobacteria and cyanobacteria hybrid PKS I/NRPS gene clusters were dominant.

From the 13 bacterial genomes possessing three or more PKS I genes, seven can be assigned to the actinobacteria, four to the cyanobacteria, one to Bacillales, and one to the pseudomonads. These results are not representative of the distribution of PKS I genes in bacteria as several bacterial species and genera are underrepresented in the current list of completely sequenced microbial genomes, whereas other bacterial groups are overrepresented. In particular, no mycobacterial genome sequence is currently available in the public database. The results of this survey are nevertheless in agreement with the number of metabolites that have been reported for the individual bacterial groups.

### Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Genomes</th>
<th>Number of Positive Genomes</th>
<th>Number of PKS I Proteins</th>
<th>Number of Modules</th>
<th>Range of Modules/Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actinobacteria</td>
<td>13</td>
<td>10</td>
<td>98</td>
<td>132</td>
<td>1–5</td>
</tr>
<tr>
<td>Streptomycetes</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>55</td>
<td>1–5</td>
</tr>
<tr>
<td>Mycobacteria</td>
<td>5</td>
<td>5</td>
<td>70</td>
<td>74</td>
<td>1–2</td>
</tr>
<tr>
<td>Others</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Chlamydia</td>
<td>7</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>11</td>
<td>5</td>
<td>34</td>
<td>35</td>
<td>1–2</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>39</td>
<td>2</td>
<td>6</td>
<td>13</td>
<td>1–3</td>
</tr>
<tr>
<td>Proteobacteria (total)</td>
<td>60</td>
<td>13</td>
<td>37</td>
<td>60</td>
<td>1–6</td>
</tr>
<tr>
<td>α-Proteobacteria</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>β-Proteobacteria</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>γ-Proteobacteria</td>
<td>33</td>
<td>5</td>
<td>11</td>
<td>12</td>
<td>1–2</td>
</tr>
<tr>
<td>δ-Proteobacteria</td>
<td>2</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ε-Proteobacteria</td>
<td>5</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Spirochaetales</td>
<td>4</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Others</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Sum</td>
<td>141</td>
<td>30</td>
<td>177</td>
<td>242</td>
<td>—</td>
</tr>
</tbody>
</table>
An overview about the compounds that can be related to modular PKS I encoded by complete bacterial genomes is shown in Table 2. In S. avermitilis, 16 of the 22 proteins can be assigned to known polyketide structures, but only 1 out of 22 proteins encoded by the genome of the cyanobacterium N. punctiforme can be related to a secondary metabolite. It is therefore unknown, how many of the genes detected in this survey are really functional and how many of the corresponding enzymes are only induced under specific environmental conditions.

PKS I and the Genome Size of Bacteria

We have tested whether there is a correlation between the genome size of the bacteria and the presence of PKS I genes. Figure 3 shows the number of single PKS I modules encoded by the individual bacterial genomes in relation to the genome size. These two values showed a statistically significant correlation. Small bacterial genomes with less than 2 Mbp generally lack these genes. From the eight bacterial genomes exceeding a genome size of 7 Mbp only one strain, namely, Bradyrhizobium japonicum lacks PKS I genes. Thus, a trend toward the maintenance, duplication, and diversification of PKS I genes in bacterial genomes of larger size and the absence of those genes from reduced bacterial genomes is indicated. However, there are a number of medium size genomes, in particular those from four mycobacteria that encode a high number of PKS I modules. These pathogenic bacteria have reduced some of their metabolic pathways during coevolution with their host cells while maintaining the secondary metabolite genes (Vissa and Brennan 2001). Part of the mycobacterial PKS I genes are involved in the synthesis of specific cell wall lipids that play an essential role in host cell–pathogen interactions (Brennan and Nikaido 1995). This coincides with the fact that M. tuberculosis possesses about 250 distinct enzymes involved in lipid metabolism compared to only 50 in E. coli (Cole et al. 1998). In the majority of bacteria the biological function and the putative ecological role of polyketides are not well understood. Most of the bacteria producing multiple PKS metabolites have been rarely investigated in the context of their natural ecosystems, and
no final conclusion can be drawn about the percentage of metabolites exhibiting a true “biological function.” It is however striking that most bacteria encoding three or more PKS I proteins show complex morphological differentiation pattern, as known from actinobacteria, myxobacteria, and heterocyst-forming cyanobacteria (Meeks et al. 2002; Gehring et al. 2004).

Even though a minority of bacterial strains has maintained and expanded the ability to produce PKS I the list of individual strains includes members from all major bacterial groups. This raises the question whether there are differences between these groups in acquiring, retaining, and expanding their PKS stock. We have therefore initiated a phylogenetic analysis of these enzymes.

Phylogenetic Analysis of AT and KS Domains

A total of 139 AT domains derived from the complete genome analysis were subjected to a phylogenetic study (for details see Materials and Methods). Furthermore, AT sequences from four myxobacterial PKS pathways and from a pseudomonal pathway were included as these groups of proteobacteria were clearly underrepresented in the complete genome survey considering the high number of polyketides that have been described.

A Bayesian analysis of bacterial PKS I AT domains revealed two major clades (fig. 4). Similar topologies were obtained using ML, maximum parsimony, and distance methods. One distinct clade comprising groups A1–A4 contains all AT domains presumably activating malonyl-CoA (based on characterization or prediction) and a few domains with unpredictable substrate. A second clade consists of domains presumably activating methylmalonyl-CoA or rare substrates (groups A6–A8) and of one group of domains that are known or predicted to activate malonyl-CoA (A5). The same topology in the phylogenetic tree was obtained after the extraction of those residues from the alignment that form part of the active center of the domains (corresponding to residues Q11, Q63, G90, H91, L93, G94, R117, S200, H201, N231, Q250, and V255 of E. coli FabD, data not shown). Thus, the distinct subclusters in the phylogenetic tree do not only reflect a functional specialization of the AT domains but also the evolutionary relationships between the domains. It can be assumed that primary AT domains of bacterial PKS I activated malonyl-CoA as a substrate and evolved from the malonyl-CoA activating ancestor protein involved in fatty acid biosynthesis. Gene duplications and subsequent functional specialization toward novel substrates may have led to the evolution of AT domains clustering in the second clade of the tree (A5–A8). The similarity between actinobacterial sequences from the first clade of the tree to those of the second clade of the tree does not exceed 50%, whereas the actinobacterial sequences within both clades show at least 70% similarity. Probably, the “invention” of AT domains using substrates different from malonyl-CoA occurred only once in the evolution of modular PKS systems.

Apart from the different substrate specificities, none of the eight subgroups can be related to an obvious functional divergence of the corresponding AT domains. AT domains were found to cluster independently from the domain composition of a PKS I module, e.g., the presence or absence of a KR domain. Furthermore, the presence of NRPS modules at the donor or acceptor side has no impact on the position within the phylogenetic tree. We therefore conclude that the subgroups of the tree mostly reflect the evolutionary relationships between the AT domains that are not superimposed by functional differences. The genealogy of AT domains within and between the major bacterial groups will be discussed below.

Fig. 3.—Correlation between genome size and the number of PKS I modules encoded by 141 bacterial genome sequences. Filled diamonds represent genomes missing PKS I genes. Empty characters represent bacterial strains possessing PKS I genes. Actinobacterial strains are shown as diamonds, cyanobacterial strains as triangles, and all other strains as quadrates. Strain names are shown for strains encoding three or more PKS I. Test for a non-parametric Spearman correlation gave the correlation coefficient of $r = 0.476$ (95% confidence interval 0.33–0.60) and a $P$ value of $P < 0.0001$. 

Fig. 4.—Phylogeny of AT domains of bacterial type I PKS, inferred by Bayesian estimation. Numbers above branches indicate posterior clade probability values. Branch length indicates number of inferred amino acid changes per position. Branches are colored according to their affiliation to a bacterial group as shown in the color code. AT domains predicted to use malonyl-CoA are highlighted green, those predicted to use methylmalonyl-CoA or rare substrates orange. Tips of the tree give the names of the organisms, proteins (if annotated in the database), module number, and substrate specificities (H, malonyl-CoA; C, methylmalonyl-CoA; MB, methylbutyryl-CoA; X, unclear). Biochemically characterized AT domains are indicated with black dots. AT domains with theoretically predicted substrate specificities are indicated with diamonds. Boxes with names of polyketide compounds relate to subgroups exclusively or predominantly involved in the biosynthesis of that compound. Numbers in the side bar indicate group numbers used in the text. Abbreviations of biochemically characterized PKS I are listed in table 2.
The phylogenetic relationships of the KS domains of bacterial PKS I are provided as supplementary material (Supplementary Material online). The phylogenetic analysis of PKS I is based on the identification of homologous sequences within a genome. In contrast, detection of HGT is more cumbersome and prone to uncertainty (among others Ragan 2001). The best way to analyze HGT is to use a combination of different methods. In our analysis we used anomalous distribution of genes, phylogenetic tree incongruities, and atypical gene compositions as an indication of HGT.

Effect of Duplications and Gene Transfer on the Evolution of Bacterial PKS I

Detection of gene duplication is usually based on the identification of homologous sequences within a genome. In contrast, detection of HGT is much more cumbersome and prone to uncertainty (among others Ragan 2001). The phylogenetic tree of AT domains (fig. 4) reveals a varying impact of gene duplications and potential HGT events on the evolution of PKS I in the different groups of bacteria. In order to assess the influence of these different factors more systematically, the number of single-gene duplications and possible HGT events were assessed individually for the different bacterial groups (for details see Materials and Methods). The results based on the phylogenetic tree of AT domains are summarized in table 3. Similar results were obtained for the phylogenetic tree of KS domains (see Supplementary Material online). From this analysis bacteria possessing PKS I genes can be classified into three groups: a first group in which most or all PKS I genes stem from common ancestors and have evolved by gene duplication events; a second group including bacteria that have acquired PKS I genes secondarily by HGT without further advancement by gene duplications; and a third group in which PKS I genes may have evolved by a combination of HGT and gene duplication events. Actinobacteria and cyanobacteria fall into the first category. The fact that most or all lineages of PKS I genes have evolved from common ancestors within these bacterial groups does not exclude HGT events between single strains of actinobacteria.
and cyanobacteria, respectively. However, these intragenomic HGT events were not assessed by the phylogenetic approach used in this study. In cyanobacteria, one sequence shows clear indications for an HGT event. This sequence is further exceptional among cyanobacteria as it represents the only cyanobacterial AT domain predicted to activate methylymalonyl-CoA. The sequence clusters between a number of myxobacterial sequences in the subbranch A6. Nevertheless, cyanobacteria show a much stronger impact of internal sequence duplications and were therefore classified into the first category. The second category of sequences derived from HGT events without further gene duplications was detected in a few genomes of α-, β-, and γ-proteobacteria. The distribution of these sequences in the phylogenetic tree clearly indicates an ancestry from other bacterial groups rather than a common origin. Finally, myxobacteria, which belong to the δ-group of proteobacteria, show many duplication events as well as substantial gene import by HGT and thus fall into the third category.

Even though the irregular distribution of sequences can be taken as a first indication of HGT events, additional evidence is required to finally prove this theory. We have therefore analyzed the GC contents of those nucleotide sequences that encode AT domains suspected to activate methylymalonyl-CoA. The sequence clusters between a number of myxobacterial sequences in the subbranch A6. Nevertheless, cyanobacteria show a much stronger impact of internal sequence duplications and were therefore classified into the first category. The second category of sequences derived from HGT events without further gene duplications was detected in a few genomes of α-, β-, and γ-proteobacteria. The distribution of these sequences in the phylogenetic tree clearly indicates an ancestry from other bacterial groups rather than a common origin. Finally, myxobacteria, which belong to the δ-group of proteobacteria, show many duplication events as well as substantial gene import by HGT and thus fall into the third category.

In α-, β-, and γ-proteobacteria most of the PKS I sequences are located either on pathogenicity islands or plasmids that are generally accepted to be the result of HGT (Dobrindt et al. 2004). In particular, striking are the positions of two AT domains from Pseudomonas syringae (1 and 2, fig. 2) that cluster in close proximity of actinobacterial sequences in groups A2 and A7. Both proteins are part of the coronofacic acid biosynthesis complex that is involved in the biosynthesis of the phytotoxin coronatine (Bender, Alarcon-Chaidez, and Gross 1999). The corresponding nucleotide sequences show a GC content of 68%, a value that is rather similar to the GC content of streptomycetes but significantly deviating from the average GC content of pseudomonades that is 58%. Thus, in this case an HGT from streptomycetes to P. syringae is very likely. A third AT domain detected in the genome of P. syringae (Irpl1) clusters in the direct neighborhood of a gene cluster that corresponds to the Irp11 region from Yersinia pestis that is involved in the biosynthesis of the iron chelator yersiniabactin. Irp1 consists of an NRPS module and a PKS module (Miller et al. 2002). As seen for coronatine biosynthesis genes, the corresponding nucleotide sequences in P. syringae and Y. pestis significantly exceed the average GC content with 65% and 61%, respectively. Thus, the yersiniabactin biosynthesis gene cluster may originate from a host bacterium with a high GC content, e.g., an actinobacterium and was transferred to Y. pestis and P. syringae via HGT. Altogether, in the different groups of proteobacteria we can find three kinds of evidence for HGT of PKS I genes: anomalous distribution of these genes, incongruities between phylogenetic trees, and deviating GC contents of PKS I genes.

**Multiple PKS I Gene Clusters as Evolutionary Traits to Increase Metabolic Diversity**

In the previous sections we have discussed that multiple duplication events are the basis of the evolution of modular PKS systems. In this context the question arises to which extent modules are duplicated and whether modification of duplicated units occurs by means of recombinational exchange or the loss of domains. Such recombination processes generally seem to play an important role in creating variability within bacterial genomes (Smith 1991).

When we look at the domain structure of modules belonging to the same biosynthesis cluster it becomes clear that duplication alone is not sufficient to reconstruct their formation. As an example, the protein AveA4, which is part of the avermectin biosynthesis cluster of S. avermitilis, comprises three complete PKS modules showing the domain structures KS-AT-KR-ACP, KS-AT-ACP, and KS-AT-DH-KR-ACP, respectively. When we assume that the protein has evolved by duplications these domain structures can only be explained by subsequent loss or acquisition of KR and DH domains.

More information about the impact of recombination events on the evolution of modular PKS enzymes comes from the analysis of the corresponding AT and KS domains within the phylogenetic trees. The oligomycin biosynthesis protein OlnA6 from S. avermitilis provides an example for a three-modular PKS I that could have evolved by gene duplication, only. The modules show the same domain organization (KS-AT-KR-ACP), they all use the same substrate, and both the KS domains and the AT domains cluster very...
near to each other in the same subgroup of the respective phylogenetic trees (see fig. 4 and Supplementary Material online). This is supported by the fact that the respective DNA sequences are also very similar, even in the interdomain regions, which normally show a relatively high degree of variability (data not shown). Recombination events could explain the different substrate specificities of AT domains within single biosynthesis proteins or pathways. As discussed above, malonyl and methylmalonyl-CoA using AT domains show extensive amino acid differences, and therefore an independent evolution of a single methylmalonyl-CoA activating AT domain is very unlikely. Instead, it can be assumed that in those proteins containing both types of AT domains the duplication of a KS-AT unit was followed by an exchange of one of the AT domains by means of recombination. The different examples show that intra- and intergenomic recombination may have contributed to the evolution of PKS I in bacteria. To address this question more deeply it would be necessary to analyze the nucleotide sequences of all domain types and interdomain regions within the single bacterial genomes.

The generation of a potent biomolecular activity can be considered as a rare event in evolution, taking into account that such an activity is based on very specific interactions between molecules (Jones and Firn 1991). Likewise, one can infer that the process of developing bioactive secondary metabolites requires a considerably long span of time. Firm and Jones (2000) proposed a unifying model for the evolution of secondary metabolites. In their model they suggest that organisms may have selected specific evolutionary traits to increase the chances to develop a compound with potent biomolecular activity. The appropriate traits should enhance the generation and retention of chemical diversity and concurrently should reduce the fitness costs. Firm and Jones propose that by changing a single enzyme component in a biosynthesis pathway where each of the enzymes exhibits broad substrate specificity an organism can create diverse new products. The diversity of modular PKS I products, however, seems to arise rather from frequent recombination events between the modules than from evolution toward broader substrate specificities. Only modular systems like bacterial PKS I provide such an extraordinary platform for recombination events. This could explain the selective advantage for bacteria possessing multiple PKS gene clusters. At first sight it seems to be a genetic burden and paradox to keep these extremely large gene clusters in the genome, some of them not even producing a compound with biological activity. However, the valuable evolutionary advantage is their effect as a “gene-saving device” (Cerda-Olmedo 1994) because the organisms have the ability to produce a large chemical diversity using a very limited number of different genes. This applies not only for molecules with an antibiotic activity like in the case of streptomycetes but also for compounds that may play a role in cell processes like signaling and communication.

Supplementary Material

Supplementary materials are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Prof. T. Börner (Humboldt University, Berlin) for critical reading of the manuscript and Dr. I. Schmitt (Field Museum, Chicago) for helpful suggestions. This work was supported by grants of the German Research Foundation (DFG-SPP1152) to E.D. and R.M.

Literature Cited


Downloaded from https://academic.oup.com/mbe/article-abstract/22/10/2027/1138202 by guest on 08 January 2019
Evolution of Secondary Metabolites in Bacteria


William Martin, Associate Editor

Accepted June 10, 2005