J. V. Lopez

Naturally mosaic operons for secondary metabolite biosynthesis: variability and putative horizontal transfer of discrete catalytic domains of the epothilone polyketide synthase locus

Abstract A putative instance of horizontal gene transfer (HGT) involving adjacent, discrete β-ketoacyl synthase (KS), acyl carrier protein (ACP) and nonribosomal peptide synthase (NRPS) domains of the epothilone Type I polyketide biosynthetic gene cluster from the myxobacterium *Sorangium cellulosom* was identified using molecular phylogenetics and sequence analyses. The specific KS domain of the module EPO B fails to cluster phylogenetically with other epothilone KS sequences present at this locus, in contrast to what is typically observed in many other Type I polyketide synthase (PKS) biosynthetic loci. Furthermore, the GC content of the *epoB* KS, *epoA* ACP and *NRPS* domains differs significantly from the base composition of other epothilone domain sequences. In addition, the putatively transferred epothilone loci are located near previously identified transposon-like sequences. Lastly, comparison with other KS loci revealed another possible case of horizontal transfer of secondary metabolite genes in the genus *Pseudomonas*. This study emphasizes the use of several lines of concordant evidence (phylogenetics, base composition, transposon sequences) to infer the evolutionary history of particular gene and enzyme sequences, and the results support the idea that genes coding for adaptive traits, e.g. defensive natural products, may be prone to transposition between divergent prokaryotic taxa and genomes.

Keywords Horizontal gene transfer · Epothilones · Polyketide · Phylogenetics · Operon

Communicated by W. Arber

J. V. Lopez
Division of Biomedical Marine Research, Harbor Branch Oceanographic Institution, 5600 US 1 North, Ft Pierce, FL 34946 USA
E-mail: Lopez@hboi.edu
Fax: +1-772-4612221

Introduction

Polyketides (PK) are secondary metabolite compounds generated by successive condensations of simple carboxylic acids. Polyketides represent an important class of "natural products"—secondary metabolites that include hundreds of medicinal and antibiotic compounds (Chadwick and Whelan 1992; Staunton and Weissman 2001). Organisms which produce polyketides include bacteria, fungi, dinoflagellates, sponges and plants, with the majority of bioactive PKs stemming from Gram-positive actinomycetes (Hopwood and Sherman 1990; Longley et al. 1993; Snyder et al. 2003).

The molecular biology of polyketide biosynthesis is one of the best characterized models among secondary metabolites (Hopwood 1997). For example, Type I polyketides include such complex compounds as the macrolides erythromycin and rifamycin (Donadio and Katz 1992; August et al. 1998), and are synthesized by large (100–10,000 kDa) multifunctional enzymes in which distinct catalytic domains are organized within "modules" that perform the sequential acyl extensions and modifications (Khosla et al. 1999). β-Ketoacyl synthase (KS), acyl carrier protein (ACP) and acyl transferase (AT) domains coordinate acyl chain elongation, while β-ketoacyl-ACP reductase (KR), dehydratase (DH) and enoylreductase (ER) components reduce the β-position of the growing polyketide (Fig. 1). Thus, modular Type I genetic loci, which probably share a common ancestry with genes for animal fatty acid synthases (FASs), can become quite large, spanning 30 to over 100 kb (Hopwood 1997).

The polyketides epothilone A and B are 16-membered macrolactones produced by the myxobacterium *Sorangium cellulosum*. The epothilones possess a methylthiazole moiety connected to a macrocycle by a short olefinic spacer, and show promising antitumor activities; they bind to paclitaxel-binding sites on β-tubulin, leading to G2-M arrest, apoptosis, cell death (very similar to Taxol and discodermolide) (Longley et al. 1993; Molnar

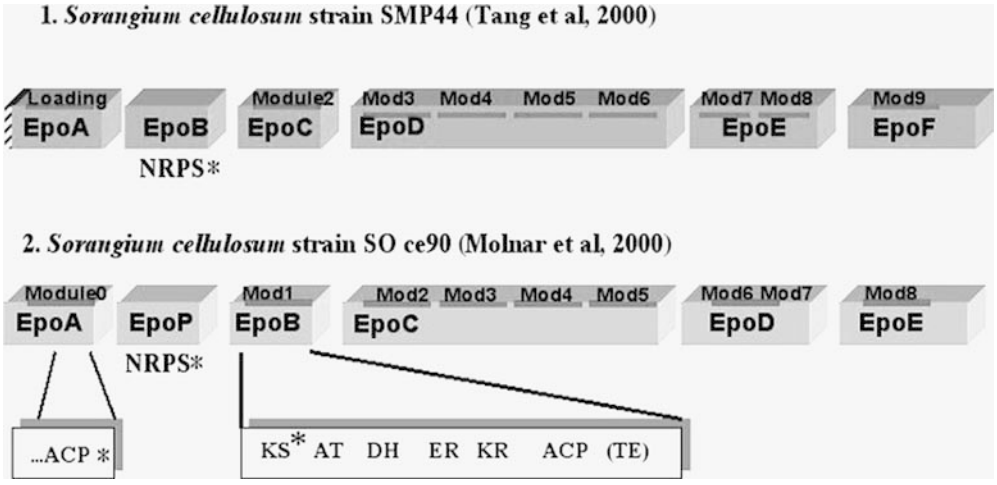

Fig. 1 Schematic diagrams of epothilone A/B loci as sequenced in two different laboratories. The epothilone locus reported by Tang et al. (2000) from *Sorangium cellusom* strain SMP44 (GenBank AF217189) is shown at the *top*; the epothilone locus from strain SO ce90 (GenBank AF210843) described by Molnar et al. (2000) is the sequence primarily used in the current analysis. The *large grey blocks* contain modules (mod) with discrete ORFs. Within each module are sequences that encode specific catalytic domains (KS, ER, DH etc.); examples are given in the *white boxes* at the *bottom*. Abbreviations for various catalytic domains are explained in the text. Although a representative domain order is shown for *epoB*, the reader is directed to the primary references of each PKS locus for the precise order and description of other epothilone modules. The *asterisks* mark the PKS regions that may have experienced a HGT. The *diagonally striped rectangle* indicates regions that show identity to transposon-like sequences, such as the extreme 5′ end of the SMP44 sequence, which includes two distinct transposon regions

et al. 2000). Epothilone also has high water solubility and works well against P-glycoprotein-expressing multiple drug resistant cell lines. The Type I PKS gene locus of the epothilone-producing *Sorangium* strain SO ce90 spans 68,750 bp, and is comprised of six ORFs, which include nine PKS modules, and one non-ribosomal peptide synthetase (*NRPS*) gene. *NRPS* loci only slightly resemble PKS loci in their modular arrangements (Marahiel et al. 1997; Du et al. 2000; Huang et al. 2001). The PKS ORF *epoC* comprises four modules and is itself 21,773 bp long.

In this paper, I describe a comparative analysis of KS regions from previously characterized Type I polyketide-producing microbes, including epothilone-producing Myxobacteria and a range of other prokaryotes, from Gram-positive actinomycetes to cyanobacteria. The KS domains show the highest degree of amino acid sequence conservation among all of the different PKS catalytic domains, and therefore retain some residual phylogenetic signal. Evidence is presented here which shows that (1) KS domains within a single operon, and to lesser extent within a species, tend to group together phylogenetically, and (2) the *epoB* KS and adjacent *epoP* (NRPS) gene sequences may have been horizontally transferred, based on the application of several criteria for horizontal gene transfer (HGT) (Koonin et al. 2001). These data have implications for the evolution of this and other PKS loci. For example, the common observation of phylogenetic grouping of domains within a locus supports a possible natural mechanism for the diversification of secondary metabolite biosynthetic loci via gene duplication, while a HGT event would support the Selfish Operon Theory (Lawrence and Roth 1996).

The impetus for this study stems from the hypothesis that the coding genes for secondary metabolite biosynthesis are relatively mobile genetic elements that can be horizontally transferred and confer adaptive value to taxonomically diverse organisms (Stone and Williams 1992; Lan and Reeves 1996; Lawrence and Roth 1996; Jain et al. 2002). This is the first in-depth study of an apparently natural horizontal transfer of a specific secondary metabolite biosynthetic gene to a related locus based on multiple criteria, such as base composition, phylogeny and the presence of transposon sequences.

Materials and methods

Sequence and phylogenetic analyses

The sequences of many Type I PKS loci were retrieved from GenBank (releases 130–135.0) for sequence analyses (Table 1). The criteria used for the selection of KS sequences were the following: (1) empirical confirmation of a polyketide metabolite, (2) availability of the full-length KS sequence, (3) interesting or unique taxonomic identity of the source organism, and (4) presence of multiple KS domains (an operon) within a single genome or species. A total of 92 distinct KS domains were extracted from these PKS loci based on the GenBank annotation for each locus; the analyzed lengths are listed in the Results section.

A potentially confusing element in the analysis of epothilone biosynthetic genes arises from the fact that two very similar epithilone operons have been fully and independently sequenced from different *Sorangium* strains—SO ce90 (Molnar et al. 2000) and SMP44 (Tang et al. 2000) (Fig. 1). Nevertheless, there exists 98.4% sequence identity between the epothilone loci of these two strains, and thus, apart from their nomenclature, both can be considered to be effectively identical for the purposes of the present study. For simplicity, the current analysis and subsequent citations focus primarily on the SO ce90 sequence reported by

Table 1 PKS sequences retrieved from GenBank for the present analysis

Genetic locus	PK metabolite[a]	Source organism	GenBank Accession Nos.
	Microcystin	*Anabaena sp. 90*	AAO62584.1, AAO62584
lovastatin	Lovastatin	*Aspergillus terreus*	AF141925
	Putative	*Bacillus subtilis*	CAB13603
Barbamide E	Barbamide	*Lyngbya*	AAN32979
McyD	Microcystin	*Microcystis aeruginosa* PCC7806	AAF00959
ppsB-D	Phenolpthiocerol	*Mycobacterium bovis*	CAD96644.1 - CAD96646.1
pksC, D, E	Putative	*Mycobacterium leprae*	S73013, S73021, NP_302534
ppsA	Putative	*Mycobacterium tuberculosis H37Rv*	Z74697, CAA98988
NdaC, NdaD	Putative	*Nodularia spumigena*	AAO64405
	Putative	*Nostoc (Anabaena sp. strain PCC 712)*	BAB78014.1
	Putative	*Nostoc punctiformis*	ZP_00110274.1
NosB	Nostopeptolide A	*Nostoc* sp. GSV224	AAF15892
MSAS	6-Methylsalicylic acid	*Penicillium patulum*	X55776
	Microcystin	*Planktothrix agardhii*	CAD29793.1
cfa6, cfa7	Coronafacic acid	*Pseudomonas syringae*	AAD03047.1, AAD03048.1
DEBS-6 (6-deoxy-erythronolide B synthase)	Erythromycin	*Saccharopolyspora erythraea*	M63676. M63677
Fix-23	Fatty acid synthase (FAS)	*Sinorhizobium meliloti*	X64131
epoA-F	Epothilone	*Sorangium cellusom SMP44*	AF217189
epoA-E	Epothilone	*Sorangium cellusom SO ce90*	AF210843
Mxa C- F	Myxalamid	*Stigmatella aurantiaca*	AF319998_9, AF319998_9
MtaB	Myxothiazol	*Stigmatella aurantiaca*	AF188287
MtaD	Myxothiazol	*Stigmatella aurantiaca*	AAF19812
	Putative	*Streptomyces antibioticus*	S43048. AAA19695.1
	Oleandomycin	*Streptomyces antibioticus*	Q07017
olmA 1–7	Oligomycin	*Streptomyces avermitilis*	AB070940
	Putative	*Streptomyces avermitilis*	AB070934 - AB070957
	Enediyne neocarzinostatin	*Streptomyces carzinostaticus*	AAM77986.1
SCO6274	Putative	*Streptomyces coelicolor A3(2)*	NP_630013.1, NP_630373.1
	Putative	*Streptomyces hygroscopicus*	T03224, T03222
rapA	Rapamycin	*Streptomyces hygroscopicus*	X86780
fkbA	FK506	*Streptomyces sp.*	1781344, CAA71463
	Putative	*Streptomyces venezuelae*	T17410, T17409
blmVIII	Bleomycin	*Streptomyces verticillus*	AAG02357
LnmI	Leinamycin	*Streptomycles atroolivaceus S-140*	AAN85522
PedH	Pederin	Symbiont bacterium of *Paederus fuscipes*	CAE01108.1, CAE01106.1
xabB	Albicidin	*Xanthomonas albilineans*	AAK15074, AF239749

[a]Putative indicates that no proven PK metabolite has been identified to date; these KS designations are based on sequence similarity

Molnar et al. (2000) and its putatively transferred KS domain in the *epoB* module (which is equivalent to the " *epoC* " region in strain SMP44). EPO B, encoded by the single module *epoB* (Molnar et al. 2000), encompasses only one multifunctional enzyme. Within this module, the *epoB* beta-ketoacyl synthase (KS) gene region spans 1277 bp (positions 16269–17546), coding for roughly 425 amino acids. The nomenclature used for the two independently derived epothilone locus KS sequences differs, but *epoB* -module 1 of strain SO ce90 (Molnar et al. 2000) is effectively identical to *epoC* -module2 of SMP44 (Tang et al. 2000) (see Fig. 1).

Multiple alignments were initially made with Clustal W (Thompson et al. 1994). Since KS nucleotide sequence divergences were often too large to permit reliable alignment and subsequent phylogenetic analyses, all phylogenetic reconstructions shown here were based on amino acid sequences. Alignments were re-evaluated manually using secondary-structure predictions provided via the SAINT (Structure Assignment with Instructive Transparency) bioinformatics workbench, on the DARWIN server (Gonnet and Benner 1991) at http://www.scinq.org/~darwin/Saint.html. DARWIN provides a variety of informatics tools such as patrician tree-based data structures for genomic sequence data, rapid searches, evolutionary tree reconstruction that exploits PAM distances to which are coupled variances (which provides firmer grounds for a statistical evaluation of the quality of a tree), SIAPrediction (prediction of Surface/Interior/Active site) and ParsePrediction. SAINT also permits the estimation of Ka/Ks values and hydrophobic parameters. Currently, no empirical crystal structures for the β-ketoacyl synthase (KS) domain are available for comparison and further refinement of alignments based on known secondary structures. Final sequence alignments are available upon request.

Phenetic distance matrices and trees (neighbor-joining) were constructed using PAUP * version 4.0b10 (Nei and Kumar 2000; Swofford 2001). As a secondary analysis, maximum likelihood analysis of protein sequences was performed using the TREE-PUZZLE program (http://www.tree-puzzle.de/) (Strimmer and von Haeseler 1996), using the BLOSUM 62 model for distantly related proteins (Henikoff and Henikoff 1992)

Base composition analysis

The overall base composition and the GC content of all codon positions were determined using the program GCUA (General Codon Usage Analysis; McInerney 1998). GC values for the overall sequence, plus all three codon positions, were tested for outliers. Standard errors (SE) of codon positions that were ≥2×SE higher or lower than the means of all other KS loci or ORFs in the same organism were taken as significant indicators of horizontal transfer after Lawrence (1998). A one-tailed t-test was also performed as a secondary statistical analysis (Underwood 1998).

Table 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
1 DEBS-mod1	-	0.35	0.35	0.38	0.38	0.53	0.37	0.38	0.36	0.38	0.38	0.38	0.36	0.38	0.39	0.49	0.40	0.50	0.51	0.50	0.49	0.58	0.46	0.38	0.36	0.37	0.47	0.48	0.49	0.47	0.49	0.49	0.51	0.51	0.45	0.50	0.40	0.61	0.63
2 DEBS-mod3		-	0.34	0.35	0.38	0.55	0.36	0.35	0.34	0.36	0.35	0.36	0.36	0.37	0.43	0.48	0.47	0.47	0.54	0.52	0.52	0.59	0.42	0.42	0.34	0.35	0.50	0.50	0.51	0.52	0.52	0.51	0.52	0.53	0.48	0.52	0.43	0.59	0.61
3 DEBS-mod4			-	0.34	0.36	0.54	0.35	0.35	0.36	0.36	0.34	0.36	0.34	0.36	0.35	0.47	0.44	0.46	0.51	0.50	0.50	0.56	0.44	0.39	0.36	0.37	0.47	0.46	0.50	0.48	0.49	0.48	0.49	0.48	0.46	0.50	0.43	0.62	0.62
4 olmA3-mod1				-	0.29	0.53	0.31	0.32	0.31	0.33	0.32	0.32	0.32	0.35	0.40	0.48	0.46	0.48	0.51	0.49	0.49	0.58	0.41	0.38	0.32	0.35	0.46	0.47	0.46	0.46	0.50	0.49	0.49	0.51	0.45	0.52	0.41	0.62	0.62
5 olmA3-mod2					-	0.52	0.27	0.30	0.31	0.32	0.33	0.34	0.31	0.37	0.40	0.48	0.48	0.48	0.51	0.49	0.50	0.58	0.43	0.38	0.29	0.31	0.48	0.49	0.48	0.49	0.48	0.51	0.51	0.52	0.46	0.52	0.39	0.62	0.58
6 olmA1-mod1						-	0.49	0.52	0.52	0.55	0.52	0.55	0.53	0.54	0.47	0.50	0.49	0.50	0.50	0.48	0.48	0.54	0.51	0.52	0.54	0.54	0.50	0.50	0.51	0.50	0.53	0.49	0.50	0.40	0.46	0.43	0.55	0.60	0.57
7 olmA1-mod3								0.32	0.31	0.31	0.31	0.31	0.29	0.34	0.40	0.46	0.46	0.47	0.53	0.49	0.51	0.58	0.42	0.39	0.31	0.32	0.48	0.48	0.49	0.47	0.48	0.49	0.48	0.49	0.46	0.51	0.39	0.60	0.58
8 RIF-mod6								-	0.18	0.19	0.20	0.23	0.21	0.29	0.37	0.49	0.45	0.47	0.53	0.51	0.51	0.58	0.41	0.36	0.30	0.30	0.49	0.50	0.50	0.49	0.49	0.49	0.50	0.50	0.45	0.51	0.39	0.62	0.63
9 RIF-mod8									-	0.22	0.21	0.24	0.24	0.28	0.38	0.50	0.47	0.48	0.54	0.53	0.54	0.59	0.41	0.38	0.30	0.31	0.50	0.51	0.50	0.50	0.51	0.50	0.51	0.52	0.46	0.53	0.39	0.62	0.63
10 RIF-mod10										-	0.22	0.22	0.22	0.30	0.41	0.51	0.49	0.50	0.54	0.53	0.54	0.57	0.44	0.39	0.30	0.30	0.50	0.51	0.49	0.48	0.50	0.50	0.51	0.53	0.48	0.53	0.40	0.63	0.58
11 RIF-mod7											-	0.24	0.22	0.32	0.40	0.51	0.47	0.48	0.55	0.52	0.54	0.59	0.44	0.40	0.31	0.31	0.50	0.50	0.49	0.50	0.52	0.49	0.51	0.51	0.46	0.51	0.40	0.61	0.61
12 RIF-mod5												-	0.24	0.30	0.42	0.52	0.49	0.50	0.54	0.52	0.53	0.58	0.44	0.40	0.29	0.31	0.51	0.51	0.50	0.50	0.50	0.48	0.50	0.51	0.47	0.53	0.41	0.62	0.62
13 RIF-mod3													-	0.27	0.40	0.50	0.47	0.49	0.54	0.51	0.52	0.59	0.42	0.40	0.27	0.29	0.49	0.50	0.50	0.49	0.49	0.48	0.50	0.52	0.47	0.53	0.40	0.60	0.60
14 RIF-mod4														-	0.40	0.52	0.48	0.49	0.55	0.52	0.54	0.59	0.45	0.41	0.34	0.35	0.52	0.52	0.51	0.50	0.52	0.50	0.52	0.50	0.49	0.53	0.44	0.63	0.61
15 EPO C-mod5															-	0.42	0.39	0.40	0.46	0.45	0.45	0.55	0.38	0.38	0.41	0.42	0.43	0.45	0.43	0.43	0.46	0.44	0.45	0.44	0.41	0.43	0.44	0.57	0.56
16 EPO C-mod3																-	0.22	0.22	0.43	0.41	0.42	0.58	0.47	0.47	0.51	0.53	0.50	0.52	0.51	0.50	0.49	0.46	0.47	0.49	0.46	0.51	0.48	0.59	0.57
17 EPO D-mod7																	-	0.22	0.46	0.44	0.45	0.56	0.44	0.47	0.49	0.50	0.47	0.48	0.49	0.49	0.49	0.47	0.47	0.47	0.45	0.52	0.49	0.60	0.58
18 EPO C-mod4																		-	0.45	0.44	0.43	0.58	0.47	0.47	0.47	0.49	0.46	0.47	0.48	0.47	0.47	0.43	0.47	0.46	0.43	0.49	0.51	0.60	0.58
19 EPO D-mod6																			-	0.15	0.17	0.53	0.50	0.49	0.51	0.54	0.43	0.45	0.44	0.44	0.46	0.45	0.49	0.48	0.48	0.50	0.55	0.62	0.60
20 EPO C-mod2																				-	0.17	0.52	0.49	0.48	0.49	0.51	0.43	0.44	0.43	0.43	0.46	0.44	0.48	0.46	0.47	0.49	0.53	0.60	0.57
21 EPO E-mod8																					-	0.54	0.49	0.49	0.50	0.51	0.44	0.45	0.42	0.43	0.46	0.45	0.47	0.46	0.46	0.50	0.52	0.60	0.58
22 EPO B																						-	0.56	0.57	0.59	0.60	0.52	0.55	0.54	0.55	0.56	0.55	0.56	0.55	0.54	0.55	0.61	0.65	0.64
23 S. avermitilis PKS2-mod1																							-	0.33	0.39	0.41	0.47	0.48	0.48	0.48	0.50	0.51	0.48	0.48	0.46	0.49	0.45	0.59	0.57
24 S. avermitilis PKS2-mod2																								-	0.37	0.38	0.48	0.48	0.48	0.47	0.48	0.47	0.48	0.50	0.43	0.49	0.44	0.59	0.58
25 S. avermitilis PKS5-mod1																									-	0.07	0.47	0.49	0.49	0.48	0.47	0.49	0.50	0.51	0.45	0.52	0.38	0.61	0.59
26 S. avermitilis PKS5-mod2																											0.50	0.50	0.50	0.49	0.50	0.51	0.49	0.52	0.48	0.54	0.41	0.63	0.59
27 Mycobac leprae -C																											-	0.11	0.27	0.26	0.42	0.45	0.48	0.48	0.45	0.48	0.51	0.57	0.61
28 Phenolpthiocerol -B																												-	0.26	0.26	0.43	0.44	0.50	0.48	0.45	0.47	0.52	0.58	0.61
29 Mycobac leprae -E																													-	0.09	0.44	0.43	0.50	0.48	0.45	0.48	0.53	0.59	0.61
30 Phenolpthiocerol -D																														-	0.42	0.43	0.50	0.48	0.45	0.49	0.52	0.58	0.61
31 Mycobac leprae -D																															-	0.43	0.47	0.51	0.47	0.52	0.54	0.57	0.61
32 Nostoc-1																																-	0.39	0.44	0.42	0.44	0.52	0.57	0.63
33 Microcys-mod2																																	-	0.48	0.45	0.49	0.54	0.59	0.60
34 MtaB-mod1																																		-	0.45	0.36	0.55	0.58	0.55
35 MtaB-mod2																																			-	0.41	0.48	0.57	0.57
36 MxaF																																				-	0.55	0.56	0.55
37 Pseudomonas1																																					-	0.61	0.62
38 Lovastatin																																						-	0.62
39 Fix-23																																							-

Mean Distance=0.46
(SD = 0.09)

a - Abbreviations: PKS module is abbreviated as "mod" in all cases; RIF (rifamycin), FAS (fatty acid synthase); other PKS loci abbreviations are indicated in the text or Table 1.
b - Order of sequences was determined by the ClustalW algorithm.
c - The lengths of epothilone KS sequences which were aligned and compared are shown in Table 3.

Results

Divergence of β-ketoacyl synthase (KS) domains

Due to differences in annotation and the likely presence of insertion/deletion mutations, the KS regions that were analyzed in this study span about 450–480 amino acid residues of the β-ketoacyl synthase domains of various PKSs. Surveys of the Pfam database (Bateman et al. 2002) indicated that these sequences show structural similarity to the thiolase family (pfam00108) and also to chalcone synthase. Furthermore, Psi-BLAST results indicate that the N-terminal domain contains most of the structures involved in dimer formation, as well as the active-site cysteine.

Uncorrected pairwise distances between KS amino acid sequences from different microbial taxa (and loci) analyzed in this study range widely—from 0.07 to 0.71. A representative subset of KS amino acid distances is shown in Table 2; the mean distance is 0.46 (SD = 0.09), which is relatively low compared to other PKS interdomain divergences (Donadio and Katz 1992). Intraspecific (intra-operon) KS domains show greater similarity than between-species comparisons, with the following mean pairwise distances for specific loci shown in Table 1: FK506, 0.24 (n = 3), rifamycin, 0.245) n = 21; and epothilone, 0.41 (n = 28). Nevertheless, even between different species, sequence alignments revealed the presence of highly conserved motifs in this dataset. Examples include (1) EaHGTGT, (2) DPQqR, and (3) $GP(x)_4$ dtaCSsSL (where x equals any amino acid, upper case is ≥94% invariant, and lower case designates amino acids with up to 93% consensus). Motif (3) includes the active-site cysteine of β-ketoacyl synthase, which is required for this ester linkage formation in the growing chain (Motamedi et al. 1997). Overall, the higher intra-operon sequence identities suggest a common ancestor for many KS sequences within a single biosynthetic operon, and thus imply evolution via gene duplication followed by sequence diversification.

Paralogous PKS loci

Despite the fact that many genes in bacterial genomes are arranged into operons, empirical evidence for gene duplication and divergence within bacterial operons is not commonplace. One question that remains is the extent of "paralogous" duplications of genes within a single secondary metabolite biosynthetic operon or between different PKS loci within a single genome. Thus, compelling evolutionary hypotheses, such as the role of gene conversion, positive selection (Wagner 2002) or HGT in "selfish" operons (Lawrence and Roth 1996), for the evolution of physically clustered genes—which includes PKS loci—remain to be tested.

In order to explore possible paralogous duplications of secondary metabolite biosynthetic genes within a single genome, sequence and phylogenetic analyses were also performed on confirmed and predicted PKS loci in the recently published genome sequence of *Streptomyces avermitilis* (Omura et al. 2001). Phylogenetic analysis suggests that several, but not all, discrete PKS loci in *S. avermitilis* are closely related and may have evolved via gene duplication (Fig. 2). For example, the two KS domains in the PKS5 locus yield the lowest pairwise sequence distance of 0.07 among the whole KS dataset (Table 2). Furthermore, the neighbor-joining dendrogram in Fig. 2 shows that most of the KS modules from Gram-positive/high GC actinobacteria (e.g. DEBS-6, rifamycin, and FK506) form monophyletic clades with high (95%) bootstrap support. Most of the *S. venezuelae* KS domains grouped tightly, while pederin KS loci appeared more divergent. The *Stigmatella* myxothiazol and myxalimid sequences did not show strong cohesion to each other, while other distinct non-actinobacterial clades include cyanobacteria and mycobacteria KS sequences. Interestingly, in the latter, specific phenothiopcerol KS domains from *Mycobacterium bovis* show higher degrees of identity (~90%) to their orthologues in *M. leprae* than to members in their own operon, suggesting a recent speciation event rather than a duplication.

Horizontal gene transfer within the epothilone PKS locus

The first evidence for horizontally transferred epothilone KS sequences was obtained from molecular phylogenetic analyses and comparisons of archived PKS gene and enzyme sequences from different polyketide-producing species (Fig. 2). In several phylogenetic reconstructions, EPO B KS, EPO C-mod5, both putative KS modules from *Pseudomonas syringae*, and possibly oligomycin A1-module1 consistently diverged from the majority of the KS modules in their respective operons, which typically form well supported clades in the neighbor-joining tree (Fig. 2). This is consistent with the amino acid sequence data in Table 2, which shows that EPO B is most remote from all other epothilone domains. Interestingly,

►

Fig. 2 Neighbor-joining phylogeny of 80 representative KS domains based on the mean pairwise amino acid distances shown in Table 2. In most cases each OTU (operational taxonomic unit) is named after a known KS biosynthetic locus or metabolite (abbreviations are shown in Table 1). However, if the locus is associated with only a putative polyketide (Table 1), then the OTU is designated by the genus or species of the source organism and the PKS module (mod) as indicated in the respective primary reference. The genus *Streptomyces* is abbreviated as "S.". At least 250 bootstrap replicates were performed with the final bootstrap percentages (> 50%) shown at the nodes. Nodes with no percentage were not supported and often collapsed in the final bootstrap tree. The *asterisks* indicate putative HGT domains inferred by > 1 HGT criteria. The largest clade of actinomycete Type I KS sequences is indicated. Putative KS sequences from operons with more than one KS domain are listed with their respective GenBank Accession Nos.

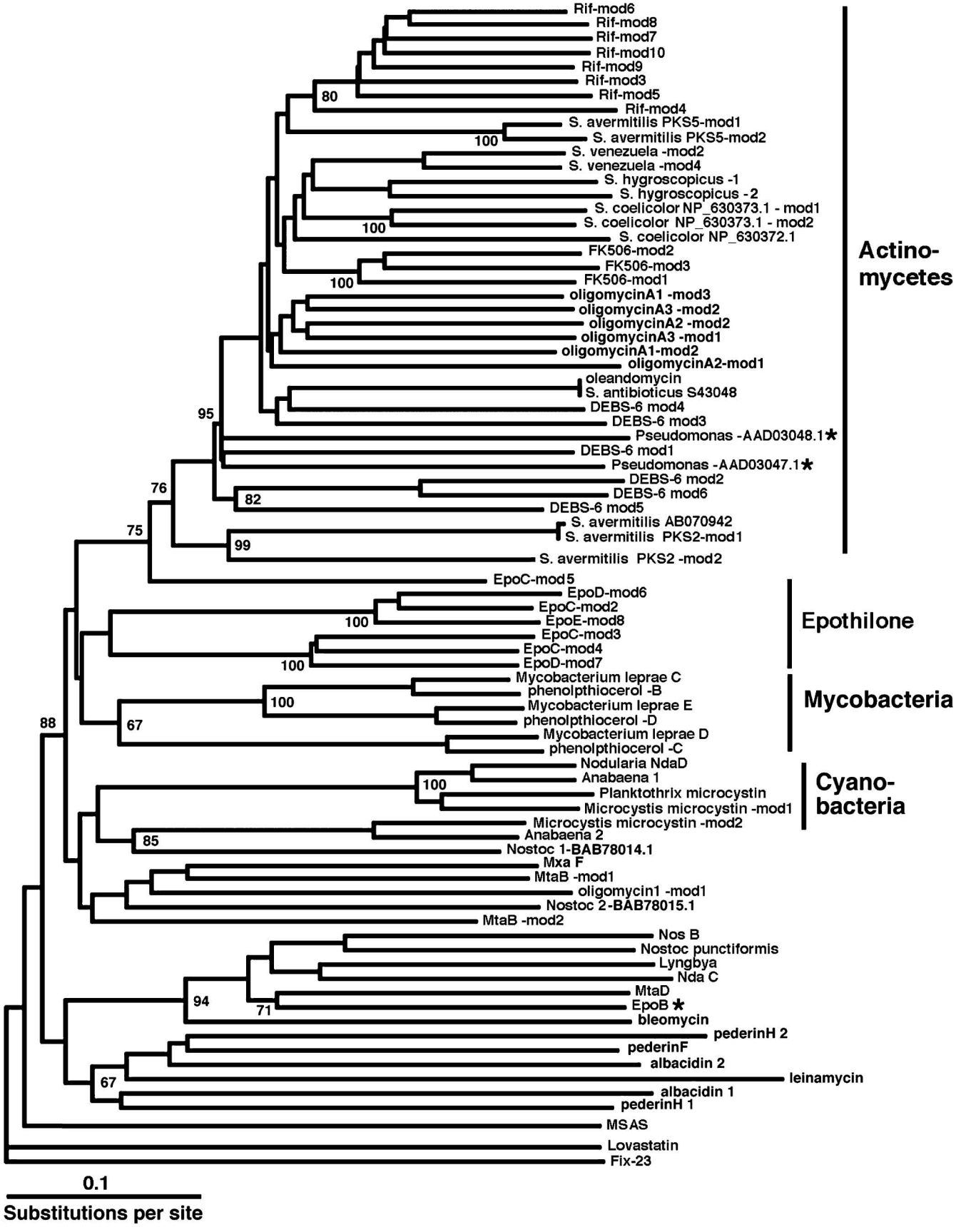
Rif-mod6
Rif-mod8
Rif-mod7
Rif-mod10
Rif-mod9
Rif-mod3
Rif-mod5
Rif-mod4
S. avermitilis PKS5-mod1
S. avermitilis PKS5-mod2
S. venezuela -mod2
S. venezuela -mod4
S. hygroscopicus -1
S. hygroscopicus -2
S. coelicolor NP_630373.1 - mod1
S. coelicolor NP_630373.1 - mod2
S. coelicolor NP_630372.1
FK506-mod2
FK506-mod3
FK506-mod1
oligomycinA1 -mod3
oligomycinA3 -mod2
oligomycinA2 -mod2
oligomycinA3 -mod1
oligomycinA1-mod2
oligomycinA2-mod1
oleandomycin
S. antibioticus S43048
DEBS-6 mod4
DEBS-6 mod3
Pseudomonas -AAD03048.1*
DEBS-6 mod1
Pseudomonas -AAD03047.1*
DEBS-6 mod2
DEBS-6 mod6
DEBS-6 mod5
S. avermitilis AB070942
S. avermitilis PKS2-mod1
S. avermitilis PKS2 -mod2
EpoC-mod5
EpoD-mod6
EpoC-mod2
EpoE-mod8
EpoC-mod3
EpoC-mod4
EpoD-mod7
Mycobacterium leprae C
phenolpthiocerol -B
Mycobacterium leprae E
phenolpthiocerol -D
Mycobacterium leprae D
phenolpthiocerol -C
Nodularia NdaD
Anabaena 1
Planktothrix microcystin
Microcystis microcystin -mod1
Microcystis microcystin -mod2
Anabaena 2
Nostoc 1-BAB78014.1
Mxa F
MtaB -mod1
oligomycin1 -mod1
Nostoc 2-BAB78015.1
MtaB -mod2
Nos B
Nostoc punctiformis
Lyngbya
Nda C
MtaD
EpoB *
bleomycin
pederinH 2
pederinF
albacidin 2
leinamycin
albacidin 1
pederinH 1
MSAS
Lovastatin
Fix-23
80
100
100
100
95
76
82
75
99
100
100
100
88
67
85
94
71
67
Actino-
mycetes
Epothilone
Mycobacteria
Cyano-
bacteria
0.1
Substitutions per site

Table 2 also shows that the epothilone locus has one of the widest ranges of pairwise distances among KS domains, varying from 0.15 to 0.58. In a maximum likelihood reconstruction using TREE-PUZZLE (data not shown), the majority of epothilone KS domains split into two clades consisting of (1) EPOC-mod3 - EPOD-mod7 - EPOC-mod4, and (2) EPOD-mod6 - EPOC-mod2 - EPOE-mod8, with 83% and 70% bootstrap support, respectively. The latter three KS domains are the most 5′ within their respective modules. Here again, EPO B never clustered with other EPO KS counterparts in either distance or maximum likelihood reconstructions.

Initial standard and psi-BLAST queries yielded the highest identities of EPO B KS-like sequences to the KS domains of the cyanobacterium *Nostoc punctiforme* (0.58) and the myxobacterium *Stigmatella aurantiaca* (0.62) (Beyer et al. 1999). In contrast, other EPO KS domains always showed their highest BLAST matches to other epothilone operon domains. The alliance of EPO B KS with other mixed or "hybrid" NRPS-KS sequences is discussed below.

Proximity of transposon-like sequences to KS domain loci

More evidence for HGT stems directly from the primary DNA and amino acid sequences of the epothilone (EPO A—E) locus itself. Tang et al. (2000) reported the presence of "transposon-like" sequences in their published sequence from *Sorangium* strain SMP44, but did not attribute any specific function to these or discuss possible transfer mechanisms involving the epothilone locus. GAP analyses with the GCG program package (Genetics Computer Group, Madison, Wis.) verifies that the first 1579 bp of the SMP44 epothilone locus (Tang et al. 2000) are not included in the published SO ce90 sequence (Molnar et al. 2000). However, the SMP44 locus actually has two distinct transposon-like reading frames of 992 and 512 bp, respectively, upstream of the first SMP44 epothilone loading module (Fig. 1). In BLAST queries (Altschul et al. 1997), the upstream SMP44 epothilone transposon sequence shows about 31% identity to IS*21* insertion element families (NP_535340.1) of *Escherichia coli* and *Agrobacterium tumefaciens*, followed by a *Pseudomonas aeruginosa* transposase (GenBank CAA32898.2). A putative *Streptomyces coelicolor* transposon sequence also match. The second downstream SMP44 transposon shows 37% identity to a *Methanosarcina mazei Goe1* transposon sequence (NP_634709.1). It is not known whether these sequences transposed simultaneously or sequentially.

GC content of KS and NRPS domains

As a third level of support for a possible HGT of the *epoB* KS and *NRPS* domains, the GC content at various codon positions in the pertinent reading frames was assessed (Table 3). The third codon position tends to be the most prone to mutational bias due its predominantly neutral characteristics (Sueoka 1992). Myxobacteria are known to have relatively high genomic GC contents, on the order of 0.70–0.72. Among all epothilone KS domains, GC content at the first and second codon positions of the putatively horizontally transferred *epoB* KS domain differed by >13 and 4×SE, respectively, and had a GC content that was 12×SE below the mean (0.70) for total GC content among all epothilone KS domains. This bias persisted after comparison with other catalytic domains of the same *epoB* module (Table 3). Interestingly, no GC bias was evident when the complete ORFs of each module (*epoA–E*) were analyzed as a whole (Fig. 1), suggesting the possibility that only individual domains (e.g. KS) or small portions of an epothilone PKS module may have been horizontally transferred. The divergent *epoC* module 5 also did not show a GC bias. A standard one-tailed t test supported these results, indicating significant (97% confidence level) GC bias at the first and second (but not the third) codon position, and in the total sequence, for *epoB* KS.

When GC content was assessed in genes adjacent to *epoB* in order to detect the possible boundaries of the putative HGT, only the epothilone *epoP* (NRPS) gene and the ACP domain of *epoA* domain showed significant departures from the mean GC content (Table 3). The *epoP* gene showed another large GC deviation, being 7×SE below the mean for the epothilone operon, and >2 and >4×SE below the mean at the first and second positions, respectively (Table 3). Changes in GC content at all positions in *epoP* GC were significant at $P < 0.01$. Since the *epoA* APC- *epoP-epoB* KS regions all lie adjacent to one another, it is possible that a single HGT could have transferred the genes as a unit.

In addition, Table 3 also includes ancillary GC analyses of *Pseudomonas*, *S. avermitilis, Nostoc*, and *Stigmatella* Mta PKS loci, because they also exhibited unusual phylogenetic placements in the neighbor-joining tree. Of these loci, only *Pseudomonas* KS domains displayed unusual GC contents, with both loci having >13×SE (and $P < 0.01$) above the mean GC content of other randomly chosen *Pseudomonas* loci, supporting another possible case of HGT.

Discussion

Several reviews (Eisen 2000; Koonin et al. 2001) have outlined many of the major criteria for determining whether a gene sequence has been horizontally transferred. Such criteria include (1) unexpected ranking of sequence similarity among homologs; (2) unexpected phylogenetic tree topology; (3) unusual phyletic patterns (determined from clusters of orthologous groups—COGs); (4) conservation of gene order between distant taxa (operons); and (5) anomalous nucleotide composition.

Table 3 GC contents of various KS, NRPS and non-PKS loci based on codon position

Gene/product	Length (bp)	Product length (aa)	GC content (%) [a]	GC content (%) [a]		
			Total	1st Pos	2nd Pos	3rd Pos
Epothilone KS regions						
EpoA	4260	1420	69.3	72.8	53.1	82.2
EpoB	5496	1832	70.4	73.4	52.6	85.2
EpoD	11394	3798	69.7	73.9	51.8	83.2
EpoE	7317	2439	69.4	73.8	51.9	82.3
EpoC	19695	6565	71.3	75.3	53.3	85.3
Mean			70.0	73.8	52.5	83.6
SD			0.8	0.9	0.7	1.5
SE			0.4	0.4	0.3	0.7
epoB-KS-mod1	1275	425	65.4**	62.6**	49.9**	83.8
epoC-KS-mod2	1257	419	70.1	70.9	54.4	85.0
epoC-KS-mod3	1278	426	71.1	74.2	51.4	87.6
epoC-KS-mod4	1278	426	69.4	72.5	51.2	84.5
epoC-KS-mod5	1269	423	68.8	69.3	51.5	85.6
epoD-KS-mod6	1260	420	71.0	70.7	54.1	88.1
epoD-KS-mod7	1275	425	71.2	73.9	52.0	87.8
epoE-KS-mod8	1257	419	69.3	71.6	52.0	84.3
Mean			70.1	71.9	52.4	86.1
SD			1.0	1.8	1.3	1.7
SE			0.4	0.7	0.5	0.6
epoP (NRPS)	4230.0	1410	63.22***	67.52***	42.91***	79.22***
Other epothilone domains						
epoA-KS-mod0	1281	427	66.4	67.5	55.5	76.4
epoA-AT	966	322	71.2	74.2	52.5	87.0
epoA-ER	900	300	69.6	75.3	47.3	86.0
epoA-ACP	216	72	63.0	59.72*	40.28**	88.9
epoB-KS	1278	426	65.4**	62.6**	49.9**	83.8
epoB-AT	963	321	71.8	75.7	51.1	88.5
epoB-DH	1248	416	71.6	81.5	67.3	66.1
epoB-ACP	213	71	69.5	69.0	45.1	94.4
epoC-KS-mod2	1257	419	70.1	70.9	54.4	85.0
epoC-AT	966	322	74.1	75.2	53.1	94.1
epoC-KR-mod2	759	253	72.7	78.3	53.8	86.2
epoD-KS-mod6	1260	420	71.0	70.7	54.1	88.1
Mean			69.7	71.7	52.0	85.4
SD			3.2	6.3	6.6	7.7
SE			0.9	1.8	1.9	2.2
Stigmatella aurantiaca -myxothiozol						
Mta-B	4632	1544	67.0	71.6	52.0	77.5
Mta-D	4644	1548	65.1	67.2	45.7	82.4
Mta-E	4719	1573	67.7	70.0	47.7	85.4
Mta-F	4800	1600	65.9	62.7	48.9	86.2
Mta-G	5154	1718	66.5	69.9	45.1	84.6
Mean			66.4	68.3	47.9	83.0
SD			1.0	3.5	2.8	3.5
SE			0.4	1.6	1.2	1.6
S. verticillus						
NRPS 12 -AF210249.1	1737	579	75.8	92.9	55.6	78.9
NRPS 11 -AAG02349.1	8520	2840	73.9	92.0	50.7	78.9
bleomycin KS	5523	1841	76.8	78.98***	56.0	95.4
Mean			75.5	88.0	54.1	78.9
SD			1.5	7.8	2.9	0.0
SE			0.9	4.5	1.7	0.0
Pseudomonas syringae						
PKS1 -AAD03047.1	6198	2066	68.4***	73.3***	51.4***	80.4***
PKS2 -AAD03048.1	8193	2731	67.7***	71.7**	50.7***	80.6***
HRPL gene -AF508897.1	552	184	54.4	61.4	35.3	66.3
Fatty-acyl isomerase -AJ535703	2295	765	60.1	60.0	43.9	76.5
Pseudomonas syringae						
sigma factor -AB016413.1	552	184	55.6	64.1	35.3	67.4
Gyrase B -AB016411	612	204	54.7	56.4	35.8	72.1
recA -AJ316163.1	597	199	58.3	64.3	35.2	75.4
RNA polymerase D -AB039622.1	807	269	60.0	70.6	40.9	68.4
Epoxidase -D82818	570	190	54.0	64.2	35.3	62.6
Methyltransferase -L49178	654	218	59.9	63.8	42.7	73.4
DNA polymerase -NC_004578	1101	367	56.8	64.9	36.5	68.9

Table 3 (Cont.)

Gene/product	Length (bp)	Product length (aa)	GC content (%) [a]	GC content (%) [a]		
			Total	1st Pos	2nd Pos	3rd Pos
Helicase -NC_004578	4908	1636	53.4	57.1	37.7	65.5
Mean			56.7	62.7	37.9	69.6
SD			2.7	4.2	3.4	4.5
SE			0.8	1.3	1.1	1.4
S. avermitilis - oligomycin						
olmA1-mod1	1608	536	75.2	77.2	56.2	92.2
olmA1-mod2	1491	497	73.4	76.1	53.3	91.0
olmA1-mod3	1584	528	75.3	75.4	56.1	94.5
olmA3-mod1	1719	573	75.0	75.4	55.7	93.9
olmA3-mod2	1719	573	74.9	74.7	55.0	94.9
PKS5-mod2	1878	626	70.3	72.8	54.0	84.2
PKS5-mod1	1908	636	70.6	73.1	54.6	84.1
PKS2-mod2	1749	583	73.5	75.3	52.5	92.6
PKS2-mod1	1809	603	75.1	76.0	56.1	93.4
Terpene cyclase NC_003155	1005	335	72.2	72.2	54.0	90.5
Transcriptional activator -NC_003155	10014	3338	74.9	80.0	55.8	89.1
Mean			73.7	75.3	54.8	90.9
SD			1.9	2.2	1.2	3.8
SE			0.6	0.7	0.4	1.1
Nostoc						
NosA-peptide synthase AAF15891.2	12681	4227	43.7	55.8	37.0	38.3
NosB-PKS-AAF15892.2	3732	1244	40.6	52.3	39.2	30.3
NosC-peptide synthase -AAF17280.1	9948	3316	44.8	57.0	37.6	39.8
NosD-AAF17281.1	7350	2450	43.5	55.9	37.2	37.3
NosE- AAF17283.1	1104	368	42.9	57.3	40.5	31.0
NosG-ABC-AAF17285.1	2238	746	35.8	44.6	34.5	28.4
Mean			41.9	53.8	37.6	34.2
SD			3.3	4.8	2.1	4.8
SE			1.3	2.0	0.8	2.0

[a]Significance levels: **($p < 0.03$), ***($p < 0.01$); SE, Standard Error; SD, Standard DeviationGenBank Accession Nos. are shown for random non-PKS genes from the same organism retrieved for comparison

In addition, Brown et al. (2001) have successfully used phylogenetics with the incongruence length difference (ILD) test to infer past recombination and HGT events in bacteria. The concordance of different types of evidence pointing to a HGT event involving *epoB* KS and *NRPS* domains avoids the dependence on any single criterion, such as base composition (Liisa et al. 2001). Since nucleotide-based phylogenies may be influenced by base composition alone (Graur and Li 1997), the finding that the EPO B KS does not cluster with its EPO counterparts in amino acid-based phylogenies also reinforces the HGT hypothesis. GC content did not have an effect on the grouping of NRPS-KS sequences in the tree shown in Fig. 2.

The presence of flanking transposon-like sequences which could provide a transfer mechanism by acting as ''vectors'' (e.g. conjugative plasmids or transducing bacteriophages) was not included in the above list of criteria, but such sequences are found in the vicinity of epothilone loci. Although his estimates have been corrected since the original publication, Lawrence (1998) has documented the phenomenon of large-scale gene transfer (approximately 18%) within the *E. coli* genome, and the association of insertion elements (IS) with 68% of putatively horizontally transferred genes. The data from *E. coli* implied that IS elements may mediate the transfer of genes, as well as gene duplication and amplification of paralogs within a single genome (Romero and Palacios 1997). Moreover, the number of examples in which secondary metabolite operons are found in juxtaposition with transposon-like sequences appears to grow in proportion to the increasing number of recently completed microbial genome sequences (Bentley et al. 2002), suggesting that HGT phenomena occur widely in other bacterial species, especially those found in the natural environment. The finding of potentially transferred KS domains in *P. syringae* is consistent with recent studies of pathogenicity islands in this microorganism (Charity et al. 2003). Omura et al. (2001) reported that 7% of the *S. avermitilis* genome—one of the highest proportions among currently sequenced bacterial genomes—codes for secondary metabolite biosynthesis (including the important antibiotic avermectin and oligomycin polyketide compounds) and appears to be associated with transposon-like genes. A recent study of the pederin biosynthetic locus, which encodes another mixed polyketide-NRPS metabolite, also documents the presence of flanking transposon-like sequences (Piel 2002). It is possible that the true number and extent of transposons in secondary metabolite HGTs may never be known, since they may become non-functional due to mutations, followed by an erosion of sequence similarity (Blot 1994). Although the putatively transferred epothilone domains may not be perfectly

bounded by transposon sequences, there are several noncoding regions between genes which could once have served as recombination hotspots for mobile vectors. Moreover, the exact molecular mechanism of transfer of these biosynthetic genes has not been adequately modeled (compared to horizontal transfer of genes by transduction or on plasmids, for example; Zgur-Bertok 1999). Therefore, the present data do not allow precise delineation of the transposed sequences by either base composition analysis or ancient transposon footprints.

Alternatives to the HGT hypothesis to explain the anomalous KS domain divergences from their operon members include (1) convergent evolution (discussed below), (2) rapid rates of amino acid substitution, (3) long time periods between duplication events among different PKS paralogous loci, and (4) long branch attraction, an artifact of phylogenetic reconstructions. Preliminary analyses for positive selection using SAINT did not show elevated levels of nonsynonymous (Ka) over synonymous (Ks) substitutions among epothilone KS sequences, which leaves open the possibility that gene conversion mechanisms may be acting to homogenize adjacent modules, or that many KS domain sequences are evolving neutrally (Ohno 1970; Wagner 2002). New algorithms for detecting positive selection within shorter protein segments are being explored (Suzuki and Gojobori 1999).

Assuming that HGT is the correct explanation for the findings reported in this study, the primordial source organism or ancestor for *epoB* KS, and other hybrid NRPS-KS domains (if they are truly evolutionarily related), will continue to loom as an intriguing question that is, in principle, addressable by phylogenetic studies. The answer may depend on the availability of novel candidate PKS sequences for comparison. Despite the growing number of PKS loci in current sequence databases (>500 entries in GenBank), many more undoubtedly remain undiscovered.

NRPS loci have been compared to PKS loci because of their modular structures and similar mode of catalysis; however, at the primary sequence level there is no evidence for shared ancestry (Sosio et al. 2000; Silakowski et al. 2001; Huang et al. 2001). Molnar et al (2000) speculate that EPO B shows less identity to other epothilone KS domains primarily due its role in the condensation reaction between 2-methyl-4 carbonyl thiazole and a methyl-malonyl group after accepting the former substrate from the upstream NRPS. Interestingly, the EPO B KS active site does diverge from all other epothilone KS sequences by exhibiting a threonine in the conserved active site motif (CSTSL). Two other recent studies (Du and Shen 2001; Moffit and Neilan 2003) present evidence to show that KS domains in hybrid NRPS-PKS loci tend to group together in phylogenetic reconstructions. Although the KS phylogeny reported in this study (Fig. 2) mirrors these previous results, interpretations of a "common" ancestry for these KS sequences, which phylogenetic trees attempt to infer (Graur and Li 1997; Nei and Kumar 2000), should be treated with caution for the following reasons. (1) There are exceptions to these trees, as not all possible mixed KS-NRPS sequences (e.g. *S. avermitilis* putative PKS2, pederin, FK506, and rapamycin loci) appear in the "hybrid NRPS-PKS" clade, in violation of the premise of Molnar et al. (2000). (2) Although the formation of a discrete hybrid NRPS-KS clade reflects similar biochemical function, genetic distances among this group were still relatively high (0.31–0.45), with a mean of 0.38, and thus evidence for an unequivocal evolutionary basis for this grouping remains ambiguous. That is, the phylogeny in Fig. 2 and previous studies, does not absolutely establish a common ancestor for these KS sequences, since sequence convergence alone, due to similarity of function (condensing amino acids to the polyketide extender units), could explain the grouping, which would then be equivalent to an evolutionary artifact.

Although similar functions appear to unite the NRPS-KS sequences, functional similarity does not explain the apparent independent deviation of the GC content of *epoB* and *epoP* from that of the proximal epothilone PKS domains. This is especially interesting in the context of the Ka/Ks analysis suggesting homogenization of sequences among epothilone domains.

Lastly, although the coupling of NRPS and PKS functions may partially explain the large EPO B divergence from other epothilone domains (Haydock et al. 1995), the inclusion of EPO B sequences within the hybrid NRPS-PKS clade does not preclude the possible involvement of *epoB* (or other KS domains) in a past HGT event.

Furthermore, with phylogenetic data comprising a major cornerstone of the evidence for HGT, concomitant principles of the discipline should be applied to all interpretations (Nei and Kumar 2000). For example, concordance of multiple evidence has proven to be a reliable approach for evaluating phylogenetic hypotheses (Avise and Ball 1990), a principle inherent in the above list of criteria. Secondly, a phenetic (distance) approach, as used in this study and in the UPGMA tree of Moffit and Neilan (2003), has the limitation that an explicit evolutionary model is sometimes not used for phylogenetic reconstruction. However, distance-based reconstruction is the method of choice when sequences are highly divergent, because positional homology is difficult to determine in alignments of PKS sequences from distantly related taxa (Nei and Kumar 2000). For this reason, KS domains from the same operon were chosen for phylogenetic analysis based on the premise that they would be more closely related to each other than to orthologous KS domains from different species. This hypothesis was upheld, except in the case of the phenolpthiocerol sequences from *M. bovis*, which are closely allied with the respective *M. leprae* KS domains (only 0.08–0.10 sequence divergence).

The second goal of this study was to apply phylogenetics to identify those Type I KS domains that were significantly divergent from the members of their respective PKS operons, a finding which could suggest a

possible HGT event. Determining the ultimate ancestral origins of horizontally transferred KS domains was not one of the primary goals, though they may be inferred from tree constructions.

This study has revealed that certain secondary metabolite genes may have experienced, or are prone to, HGT, due to the selective advantage their metabolic products are likely to confer on recipient organisms, e.g. defensive and communication-related compounds. The analysis focused on the *epoB* KS/ACP and *NRPS* domains, since they provide at least three different supporting lines of evidence for HGT and are located directly adjacent to each other. The apparent transposition of at least one KS, ACP, or a paired KS-NRPS sequence, into a distantly related epothilone gene cluster can have important implications for the evolution of PKS loci, and secondary metabolite biosynthesis in general (Stone and Williams 1992; Wiener et al. 1998; Walton 2000). For example, these results support the Selfish Operon Theory, which proposes that weakly selected, nonessential (e.g. secondary metabolite) genes with a single function are clustered physically to facilitate their horizontal transfer among genomes (Lawrence and Roth 1996). Ochman et al. (2000) asserted that natural selection is the final arbiter of the successful assimilation of laterally transferred genes to their new residence. Although not directly proving this hypotheses, the present data are compatible with the above view, because they support the possible recruitment of a foreign, albeit homologous, gene sequence into a preexisting metabolic pathway. Other HGT models are also possible (Jain et al. 2002; Wagner 2002). Furthermore, although it is unusual to find two disparate NRPS and PKS biosynthetic modules in nature, this has not prevented industrial and biomedical laboratories from investing extensively in combinatorial strategies (Tsoi and Khosla 1995; Tang et al. 2000). Indeed, as combinatorial biosynthesis aims to manipulate gene and domain ''cassettes'', the evidence presented in this study reaffirms that natural evolutionary processes of recombination and transposition have been exploiting this mechanism of creating biodiversity much longer.

Acknowledgements The author is grateful for critical reading of this manuscript by Drs. Eric Brown and Amy Wright and for statistical analyses by Karen Sandell, M.S. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 9974984 to JVL, and was carried out in accordance with current laws governing genetic experimentation in the USA. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF. This manuscript is Harbor Branch Oceanographic Institution Contribution #1519.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402

August PR, Tang L, Yoon YJ, Ning S, Muller R, Yu TW, Taylor M, Hoffman D, Kim CG, Zhang X (1998) Biosynthesis of the ansamycin antibiotic rifamycin: deductions from the molecular analysis of the rif biosynthetic gene cluster of *Amycolatopsis mediterranei* S699. Chem Biol 5:69–79

Avise JC, Ball RM (1990) Principles of genealogical concordance in species concepts and biological taxonomy. In: Futuyama D, Antonovics J (eds) Oxford surveys in evolutionary biology, vol 7. Oxford University Press, Oxford, pp 45–67

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. Nucleic Acids Res 30:276–280

Bentley SD, et al (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417:141–147

Beyer S, Kunze B, Silakowski B, Muller R (1999) Metabolic diversity in myxobacteria—identification of the myxalamid and the stigmatellin biosynthetic gene cluster of *Stigmatella aurantiaca* Sg a15 and a combined polyketide-(poly) peptide gene cluster from the epothilon producing strain *Sorangium cellulosum* So ce90. Biochim Biophys Acta 1445:185–195

Blot M (1994) Transposable elements and adaptation of host bacteria. Genetica 93:5–12

Brown EW, LeClerc JE, Kotewicz ML, Cebula TA (2001) The three R's of bacterial evolution: how replication, repair, and recombination frame the origin of species. Environ Mol Mutagen 38:248–260

Chadwick DJ, Whelan J (eds) (1992) Secondary metabolites: their function and evolution (Ciba Foundation Symposium). Wiley, Chichester

Charity JC, Pak K, Delwiche CF, Hutcheson SW (2003) Novel exchangeable effector loci associated with the *Pseudomonas syringae* hrp pathogenicity island: evidence for integron-like assembly from transposed gene cassettes. Mol Plant Microbe Interact 16:495–507

Donadio S, Katz L (1992) Organization of the enzymatic domains in the multifunctional polyketide synthase involved in erythromycin formation in *Saccharopolyspora erythraea*. Gene 111:51–60

Du L, Shen B (2001) Biosynthesis of hybrid peptide-polyketide natural products. Curr Opin Drug Disc Dev 4:215–228

Du L, Sanchez C, Chen M, Edwards DJ, Shen B (2000) The biosynthetic gene cluster for the antitumor drug bleomycin from *Streptomyces verticillus* ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. Chem Biol 7:623–642

Eisen J (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr Opin Genet Dev 10:606–611

Gonnet GH, Benner SA (1991) Computational biochemistry research at ETH. (Technical Report 154, Departement Informatik, Eidgenössische Technische Hochschule, Zürich)

Graur D, Li WH (1997) Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

Haydock SF, Aparicio JF, Molnar I, Schwecke T, Khaw LE, Konig A, Marsden AF, Galloway IS, Staunton J, Leadlay PF (1995) Divergent sequence motifs correlated with the substrate specificity of (methyl) malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases. FEBS Lett 374:246–248

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Hopwood DA (1997) Genetic contributions to understanding polyketide synthases. Chem Rev 97:2465–2497

Hopwood DA, Sherman DH (1990) Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. Annu Rev Genet 24:37–66

Huang G, Zhang L, Birch RG (2001) A multifunctional polyketide-peptide synthetase essential for albicidin biosynthesis in *Xanthomonas albilineans*. Microbiology 147:631–642

Jain R, Rivera MC, Moore JE, Lake JA (2002) Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 61:489–495

Khosla C, Gokhale RS, Jacobsen JR, Cane DE (1999) Tolerance and specificity of polyketide synthases. Annu Rev Biochem 68:219–253

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742

Lan R, Reeves PR (1996) Gene transfer is a major factor in bacterial evolution. Mol Biol Evol 13:47–55

Lawrence JG (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA 95:9413–9417

Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–1860.

Liisa BK, LB, Morton, RA, Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18:404–412

Longley RE, Gunasekera SP, Faherty D, Mclane J, Dumont F (1993) Immunosuppression by discodermolide. Ann NY Acad Sci 696:94–107

Marahiel MA, Stachelhaus T, Mootz HD (1997) Modular peptide synthetases involved in nonribosomal peptide synthesis. Chem Rev 97:2651–2673

McInerney JO (1998) GCUA: General Codon Usage Analyses. Bioinform Appl Notes 14:372–373

Moffit MC, Neilan BA (2003) Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. J Mol Evol 56:446–457

Molnar I, et al (2000) The biosynthetic gene cluster for the microtubule-stablizing agents epothilones A and B from *Sorangium cellusom* So ce90. Chem Biol 7:97–109

Motamedi H, Cai S-J, Shafiee A, Elliston KO (1997) Structural organization of a multifunctional polyketide synthase involved in the biosynthesis of the macrolide immunosuppressant FK506. Eur J Biochem 244:74–80

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, Oxford

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York

Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H, Nakazawa H, Osonoe T, Kikuchi H, Shiba T, Sakaki Y, Hattori M (2001) Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci USA 98:12215–12220

Piel J (2002) A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. Proc Natl Acad Sci USA 99:14002–14007

Romero D, Palacios R (1997) Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet 31:91–111

Silakowski B, Kunze B, Muller R (2001) Multiple hybrid polyketide synthase/non-ribosomal peptide synthetase gene clusters in the myxobacterium *Stigmatella aurantiaca*. Gene 275:233–240

Snyder RV, Gibbs PDL, Palacios A, Abiy L, Dickey R, Lopez JV, Rein KS (2003) Polyketide synthase genes from marine dinoflagellates. Marine Biotechnol 5:1–12

Sosio M, Bossi E, Bianci A, Donadio S (2000) Multiple peptide synthetase gene clusters in Actinomycetes. Mol Gen Genet 264:213–221

Staunton J, Weissman KJ. (2001) Polyketide biosynthesis: a millennium review. Natural Prod Rep 18:380–416

Stone MJ, Williams DH (1992) On the evolution of functional secondary metabolites (natural products). Mol Microbiol 6:29–34

Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Sueoka N (1992) Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:95–114

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16:1315–1328

Swofford DL (2001) PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods), Version 4. Sinauer, Sunderland, Mass.

Tang L, Shah S, Chung L, Carney J, Katz L, Khosla C, Julien B (2000) Cloning and heterologous expression of the epothilone gene cluster. Science 287:640–642

Thompson JD, Higgins D, Gibson TJ (1994) CLUSTAL version W: a novel multiple sequence alignment program. Nucleic Acids Res 22:4673–4680.

Tsoi CJ, Khosla C (1995) Combinatorial biosynthesis of 'unnatural' natural products: the polyketide example. Chem Biol 2:355–362

Underwood AJ (1998) Experiments in ecology. Cambridge University Press, London

Wagner A (2002) Selection and gene duplication: a view from the genome. Genome Biol 3:1012.1–1012.3

Walton JD (2000) Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: a hypothesis. Fungal Genet Biol 30:167–171

Wiener P, Egan S, Wellington EM (1998) Evidence for transfer of antibiotic-resistance genes in soil populations of streptomycetes. Mol Ecol 7:1205–1216

Zgur-Bertok D (1999) Mechanisms of horizontal gene transfer. Folia Biol (Praha) 45:91–96