The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from *Sorangium cellulosum* So ce90

I Molnár[1], T Schupp[2], M Ono[1], RE Zirkle[1], M Milnamow[1]*, B Nowak-Thompson[1], N Engel[2], C Toupet[2], A Stratmann[2], DD Cyr[1†], J Gorlach[1†], JM Mayo[1], A Hu[1], S Goff[1]*, J Schmid[1] and JM Ligon[1]

Background: Epothilones are produced by the myxobacterium *Sorangium cellulosum* So ce90, and, like paclitaxel (Taxol®), they inhibit microtubule depolymerisation and arrest the cell cycle at the G2–M phase. They are effective against P-glycoprotein-expressing multiple-drug-resistant tumor cell lines and are more water soluble than paclitaxel. The total synthesis of epothilones has been achieved, but has not provided an economically viable alternative to fermentation. We set out to clone, sequence and analyze the gene cluster responsible for the biosynthesis of the epothilones in *S. cellulosum* So ce90.

Results: A cluster of 22 open reading frames spanning 68,750 base pairs of the *S. cellulosum* So ce90 genome has been sequenced and found to encode nine modules of a polyketide synthase (PKS), one module of a nonribosomal peptide synthetase (NRPS), a cytochrome P450, and two putative antibiotic transport proteins. Disruptions in the genes encoding the PKS abolished epothilone production. The first PKS module and the NRPS module are proposed to co-operate in forming the thiazole heterocycle of epothilone from an acetate and a cysteine by condensation, cyclodehydration and subsequent dehydrogenation. The remaining eight PKS modules are responsible for the elaboration of the rest of the epothilone carbon skeleton.

Conclusions: The overall architecture of the gene cluster responsible for epothilone biosynthesis has been determined. The availability of the cluster should facilitate the generation of designer epothilones by combinatorial biosynthesis approaches, and the heterologous expression of epothilones in surrogate microbial hosts.

Addresses: [1]Novartis Agribusiness Biotechnology Research, Inc., PO Box 12257, Research Triangle Park, NC 27709, USA. [2]Novartis Pharma AG, Core Technology Area, WSJ-506.011, CH 4002 Basel, Switzerland.

Present addresses: *Novartis Agribusiness Discovery Institute, Inc., 3115 Merryfield Row, San Diego, CA 92121, USA. †Paradigm Genetics, Inc., PO Box 14528, Research Triangle Park, NC 27709, USA.

Correspondence: István Molnár and Thomas Schupp
E-mail: istvan.molnar@nabri.novartis.com
thomas.schupp@pharma.novartis.com

Key words: C-methyltransferase, epothilones A and B, heterocyclization domain, modular type I peptide synthase, nonribosomal peptide synthase

Received: **10 November 1999**
Accepted: **25 November 1999**

Published: **5 January 2000**

Chemistry & Biology 2000, **7**:97–109

1074-5521/00/$ – see front matter

Introduction

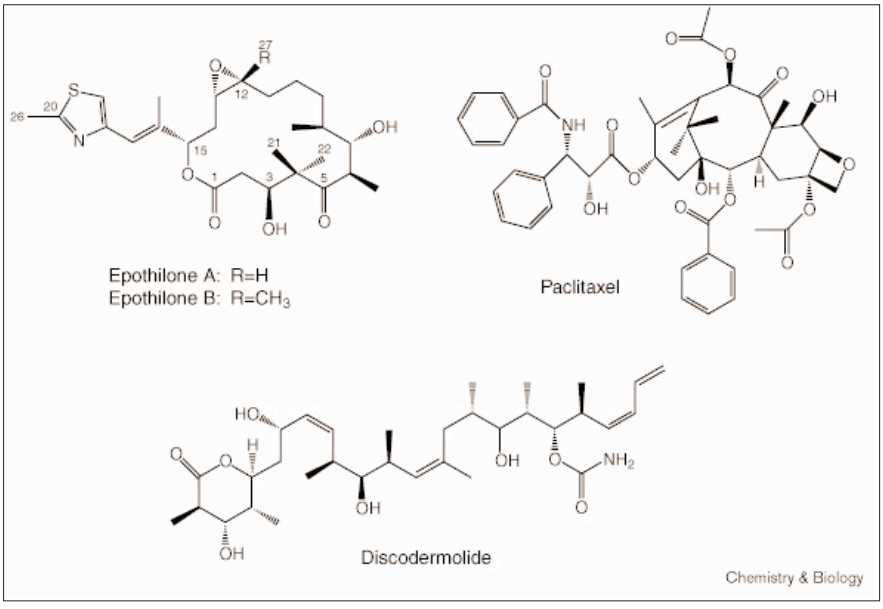
Although ~ 65% of the presently known biologically active compounds of microbial origin are produced by actinomycetes, alternative sources of molecular diversity are increasingly in demand. Along with the cyanobacteria and the basidiomycetes, strains of the gram-negative myxobacteria are an increasingly rich source of secondary metabolites that have novel basic structures and often unique modes of action [1].

The myxobacteria are ubiquitous micropredators or saprophytes in soil environments that have a gliding means of motility, complex social interactions and intriguing morphological development culminating in the production of fruiting bodies and myxospora [2]. Different strains of *Sorangium cellulosum*, from the myxobacterial order Sorangineae, were shown to produce 29 different basic structures (and many variants of these) of biologically active metabolites, amongst which only pyrrolnitrin and the jerangolids have ever been isolated from any other source [1].

Epothilones A and B (Figure 1), originally detected in 1987 as weak antifungal but cytotoxic compounds of *Sorangium cellulosum* So ce90 [3], have been rediscovered as Taxol® (paclitaxel; Figure 1) mimetics [4]. Epothilones (and discodermolide; Figure 1, [5,6]) stabilize microtubule arrays by binding to the paclitaxel-binding site on β tubulin, leading to G2–M arrest, apoptosis and cell death [7,8]. The water solubility of epothilones is significantly greater than that of paclitaxel, and they also show several thousand times higher activity than paclitaxel against P-glycoprotein-expressing multiple-drug-resistant cell lines [4,8]. All these properties warranted intensive programs for the total synthesis of epothilones A and B, as well as a growing number of structural derivatives ([9] and references therein).

Epothilones A and B are 16-membered polyketide macrolactones that have a methylthiazole moiety connected to the macrocycle by a short olefinic spacer [10]. The polyketide backbone of the molecule can be predicted to be

Figure 1



Structures of epothilones, paclitaxel and discodermolide.

synthesized by a type I modular polyketide synthase (PKS), whereas the thiazole ring might be derived from a cysteine incorporated by a nonribosomal peptide synthetase (NRPS). PKSs and NRPSs catalyze the stepwise condensation of simple metabolic building blocks. PKSs utilize coenzyme A thioester derivatives of carboxylic acids, whereas NRPSs use amino acids activated *in situ* on the enzyme as aminoacyl adenylates [11]. Both enzyme systems use carrier protein 'waystations', post-translationally modified by phosphopantheteine groups, to tether the growing chain. PKSs modify their substrates by selectively reducing the initial β carbonyl to the hydroxyl, the alkene or the alkane, and can methylate the α carbons in some cases. NRPSs add to the structural diversity of their products by epimerization, N-methylation and heterocycle formation [12]. Type I PKS and NRPS enzymes house the condensation (ketoacyl synthase for PKS versus peptidyl synthetase for NRPS), the substrate loading (acyltransferase versus adenylation), the substrate anchoring (acyl carrier versus peptidyl carrier), and the modification activities as domains on large multifunctional proteins. Both NRPSs and bacterial type I PKSs have a modular organization, whereby each successive round of elongation is catalyzed by a set of domains, used in a processive fashion [13–15]. The apparent similarities of these two pathways have been exploited to create hybrid systems of integrated PKS and NRPS modules, whereby ketoacyl or peptidyl substrates cross from one enzyme platform to another.

Although gene clusters for modular type I PKSs, NRPSs and hybrid systems have been cloned in increasing numbers in recent years (for recent reviews, see [13–15]), our knowledge of these clusters from the gram-negative myxobacteria is still limited [16–18]. Here, we report the cloning, sequencing and analysis of the gene cluster responsible for epothilone biosynthesis from *Sorangium cellulosum* So ce90. The sequenced region of 68750 base pairs (bp) encodes nine PKS modules within five multifunctional proteins, a single NRPS module encoded by an open reading frame (ORF) nested between the PKS genes, a cytochrome P450 oxidase, and 15 additional ORFs on the extremes of the sequenced region. The analysis of these genes allows the dissection of the modular logic of chain initiation with thiazole ring formation, chain elongation, termination, C-methylation, and post-PKS decoration during epothilone assembly. The availability of the biosynthetic gene cluster for epothilones should facilitate the generation of new epothilone derivatives through combinatorial biosynthetic approaches [19–21] and the heterologous expression of epothilones in host organisms that are more amenable to molecular genetics and industrial scale fermentation than the native producer *S. cellulosum* So ce90.

Results

Cloning and identification of the epothilone biosynthetic gene cluster

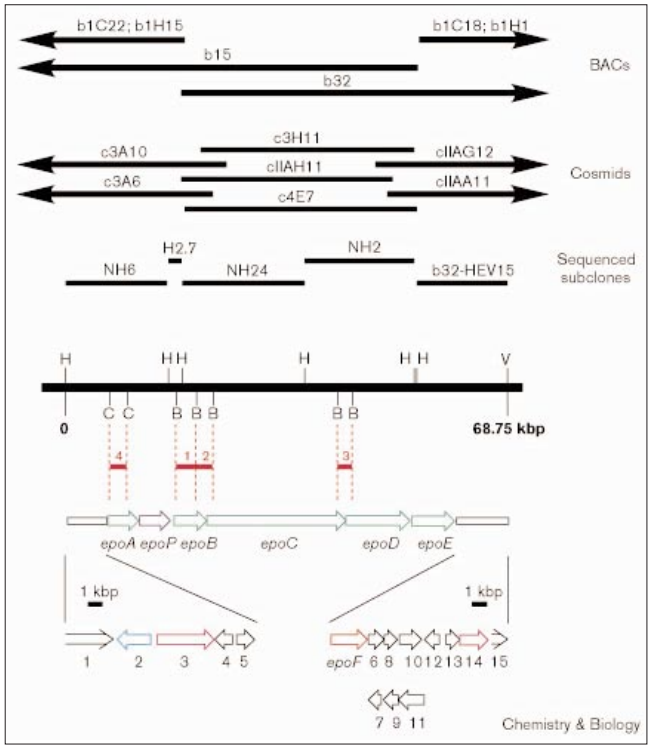
Besides epothilones, *Sorangium cellulosum* So ce90 produces spirangiens [10] and some other unidentified polyketide and/or peptide secondary metabolites. It was therefore imperative to link securely any gene clusters cloned from this organism to epothilone production in the earliest phase of the project.

A bacterial artificial chromosome (Bac) library, raised in pBelobac II [22], of *S. cellulosum* So ce90 was screened with

probes derived from ketoacyl synthase (KS) regions of the soraphen [16] and the rifamycin [23,24] cluster. Hybridizing Bac clones were sorted into five groups on the basis of their restriction patterns. *Bam*HI fragments showing sequence homologies to PKSs were subcloned into pCIB132 [16] from representative Bac clones of each group, and used for gene disruption in BCE28/2, a streptomycin-resistant derivative of *S. cellulosum* So ce90. Integration of each suicide plasmid through homologous recombination into the chromosome of BCE28/2 was verified by Southern hybridization (results not shown), and the integrants were analyzed for epothilone production. Integration of three different *Bam*HI fragments, all derived from the Bac clone b15 (fragments 1–3, Figure 2), was found to abolish epothilone production, whereas the integration of *Bam*HI fragments derived from representatives of other Bac groups had no effect on epothilone production.

The three *Bam*HI fragments were used to rescreen, at a higher stringency, the Bac library and also a cosmid library (based on ScosTriplex-II [25]) of *S. cellulosum* So ce90. An overlapping set of clones representing part of the *S. cellulosum* So ce90 chromosome (Figure 2) was compiled from hybridizing Bac and cosmid clones, and a 68750 bp region was sequenced, centered on the fragments used for disruption. To aid sequence assembly, the region was subdivided into five adjoining smaller fragments, and the fragments were sequenced separately (Figure 2).

Figure 2



Organization of the epothilone biosynthetic gene cluster. The transcriptional direction and the relative sizes of the predicted ORFs are indicated by block arrows. Numbers under the arrows refer to the ORFs whose deduced gene products are listed in Table 1. ORFs 6, 8 and 10 versus 7, 9 and 11 are alternative potential ORFs as described in the Results section. Bac and cosmid primary clones and sequenced subclones covering the 68.750-bp *Hin*dIII–*Eco*RV region (Genebank AF210843), and fragments 1–4 used for gene disruption experiments are also shown. Only the significant restriction sites are indicated for *Bam*HI (B), *Cla*I (C), and *Eco*RV (V), while all sites are shown for *Hin*dIII (H).

Organization of the epothilone biosynthetic gene cluster

Computer-assisted analysis of the sequenced region revealed 22 potential ORFs, including five PKS genes and one NRPS gene (Figure 2). The putative functions of the deduced gene products of the sequenced ORFs are listed in Table 1. CODONPREFERENCE analysis [26] of the region between base pairs 63,628 and 65,871 revealed alternative potential ORFs with plausible ribosome-binding sites in front of their start codons. As BLASTX searches of the same region did not identify any homologous sequences in the databank, we were unable to assign ORFs in this region unambiguously.

In the heart of the sequenced region are seven genes transcribed in the same direction, encoding a PKS (*epoA* and *epoB–epoE*), a NRPS (*epoP*, inserted between *epoA* and *epoB*), and a cytochrome P450 oxygenase (*epoF*, downstream of *epoE*). *epoA* and *epoP*, and also *epoB*, *epoC* and *epoD*, may be translationally coupled as judged by their overlapping stop and start codons and/or the presence of the ribosome-binding site of the downstream gene within the coding sequence of the upstream gene. No apparent transcriptional terminator could be predicted between the short intergenic regions between *epoP* and *epoB* (147 bp), *epoD* and *epoE* (15 bp), and *epoE* and *epoF* (115 bp), suggesting that all these genes form an operon of exceptional size (more than 56 kb). The promoter of this operon might lie in the unusually AT-rich stretch of 300 base pairs immediately preceding *epoA* in a 935-bp noncoding region.

Chain initiation and thiazole ring formation

Epothilones A and B contain the unusual starter unit 2-methylthiazole (Figure 1). The five-membered heterocycles of the thiazole and the oxazole series were previously shown to be formed in natural products through the cyclodehydration of cysteine (thiazole), and serine or threonine (oxazole) sidechains [27]. The presence of a PKS and a NRPS module in an intimate translational coupling at the 5′ end of the proposed epothilone synthase operon provides a plausible mechanism for the formation of this starter unit.

*epo*A encodes EPOS A, a type I modular polyketide synthase that consists of a single module, and contains a modified β-ketoacyl-synthase (malonyl-ACP decarboxylase,

Table 1

Deduced functions of ORFs in the epothilone biosynthetic gene cluster.

Polypeptide	Length	Proposed function or sequence homologies detected
EPOS A	1421	PKS
		Mod 0 KSQ* ATa ER ACP
EPOS P	1410	NRPS HC Acys OX PCP
EPOS B	1832	PKS
		Mod 1 KS ATp DH KR ACP
EPOS C	7257	PKS
		Mod 2 KS ATa KR ACP
		Mod 3 KS ATa KR ACP
		Mod 4 KS ATa DH ER KR ACP
		Mod 5 KS ATp DH ER KR ACP
EPOS D	3798	PKS
		Mod 6 KS ATp KR ACP
		Mod 7 KS ATp DH MT KR* ACP
EPOS E	2439	PKS
		Mod 8 KS ATa DH (ER) KR ACP TE
EPOS F	419	Cytochrome P450
Orf 1	> 607†	None
Orf 2	423	Hypothetical protein/DD-peptidase
Orf 3	713	Transport
Orf 4	126	None
Orf 5	149	None
Orf 6	184	None
Orf 7	145	None
Orf 8	185	None
Orf 9	146	None
Orf 10	234	None
Orf 11	288	None
Orf 12	155	None
Orf 13	156	None
Orf 14	305	Transport
Orf 15	> 134†	None

See Figure 3 for abbreviations. ATa, acyltransferase incorporating an acetate extender unit; ATp, acyltransferase incorporating a propionate extender unit; Acys, adenylation domain specific for cysteine; KSQ, modified b-ketoacyl-ACP synthase (malonyl-ACP decarboxylase). Parentheses indicate that the domains are predicted to be nonfunctional. *Domain might or might not be functional (see text).†The start codon of orf1 encoding Orf1, and the stop codon of orf15 encoding Orf15 lie outside the sequenced region.

KSQ), an acyltransferase (AT), an enoyl reductase (ER) and an acyl carrier protein domain (ACP). Although the KSQ domain at the amino terminus of the deduced EPOS A contains the two invariant histidine residues characteristic of functional KS domains [28], the KS active-site cysteine is replaced by a tyrosine. KSQ domains in loading modules of bacterial type I PKS often contain a glutamine residue at this position [29–31], similar to the so-called 'chain length factors' of type II PKS, and have been implicated in polyketide chain initiation by decarboxylation of malonyl-ACP in both systems [32]. Whether the tyrosine-containing KSQ in EPOS A or the serine-containing KSQ in the loading module of the pimaricin synthase [33] serves a similar function remains to be determined. Sequence comparisons and motif analysis [34,35] predict that the AT encoded by EPOS A would incorporate an acetate unit. The ER domain of EPOS A might or might not be functional because the highly conserved arginine in the putative ER signature motif GGVGxAAxQxA*R* (single-letter amino-acid code, where x is any amino acid) is replaced by a glutamine. EPOS A is presumed to be involved in the initiation of epothilone biosynthesis by loading an acetate unit onto the multienzyme complex that will eventually form part of the 2-methylthiazole ring (C26 and C20; Figure 3). Insertional inactivation of *epoA* by homologous recombination (*Cla*I fragment 4, Figure 2) abolished epothilone production (data not shown).

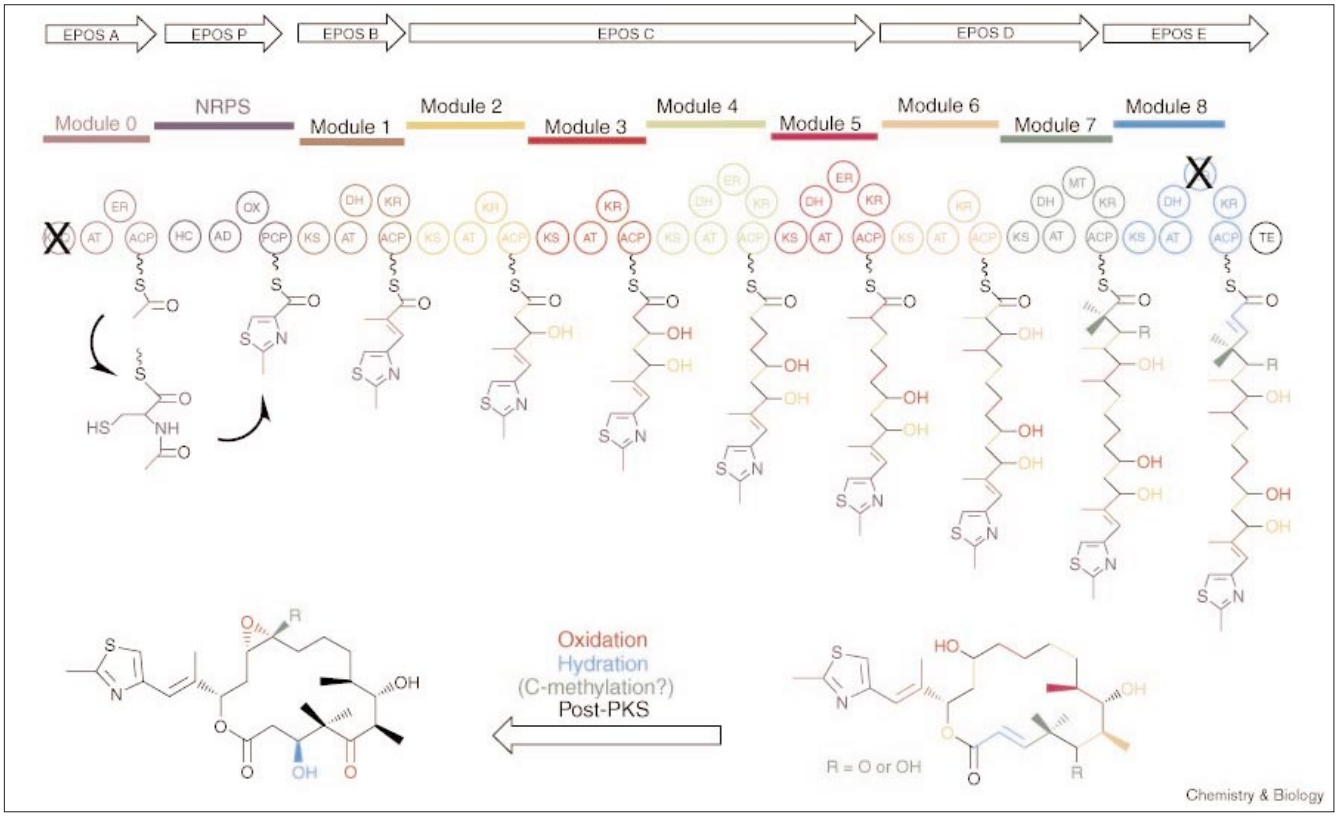
epoP codes for EPOS P, a nonribosomal peptide synthetase that consists of one module and has a heterocyclization, an adenylation, a putative oxidase and a thiolation (peptidyl carrier protein) domain, as evidenced by the presence of conserved core sequences (Table 2a, [12]). EPOS P is most similar to the heterocyclizing modules of the yersiniabactin [36], pyochelin [37] and bacitracin [38] peptide synthetases and to a recently sequenced putative peptide synthetase from *Streptomyces coelicolor* (AL023496). All of these modules, as well as EPOS P, contain a modified condensation domain (the so-called heterocyclization domain [12]) involved in the formation of the five-membered heterocyclic rings of the thiazole/oxazole series from cysteine (yersiniabactin, pyochelin, bacitracin, and also anguibactin [39] and bleomycin [40]) or serine/threonine (mycobactin/exochelin [41]).

The specificity-conferring code of NRPS adenylation domains has been solved recently [42] by comparing the sequence of the phenylalanine-activating module A (PheA) of the gramicidin S synthetase A with 160 other adenylation domains, and identifying ten amino acid residues that putatively form their amino-acid-binding pockets by superimposing the alignment on the known structure [43] of PheA. The ten amino acid substrate specificity code of EPOS P, derived by sequence alignment of the EPOS P adenylation domain with PheA, is in a complete agreement with one of the two synonymous cysteine activation signatures (Table 2b).

We propose that EPOS P is involved in the activation of cysteine by adenylation, binding of the activated cysteine as an aminoacyl-S-PCP, condensation between the enzyme-bound cysteine and the acetyl-S-ACP supplied by EPOS A, and the formation of the initial thiazoline ring by intramolecular cyclodehydration (Figure 4).

EPOS P contains a novel domain inserted between the adenylation core sequences A8 and A9. This domain shows no homology to NRPS sequences in the databanks, but displays similarity to NAD(P)H oxidases and reductases from *Pyrococcus* (AJ248285), *Methanococcus* (G64472), *Methanobacterium* (AE000801), and *Streptococcus* (AF067649)

Figure 3



Predicted domain structure of the epothilone synthase and model for epothilone formation. Each circle represents an enzymatic domain. ACP, acyl carrier protein; AD, adenylation domain; AT, acyltransferase; DH, β-hydroxyacyl-thioester dehydratase; ER, enoyl reductase; HC, modified condensation (heterocyclization) domain; KR, β-ketoacyl-ACP reductase; KS, β-ketoacyl-ACP synthase; MT, methyltransferase; OX, oxidase; PCP, thiolation (peptidyl carrier protein) domain; TE, thioesterase. Crossed out domains are predicted to be nonfunctional.

species (Figure 5a). This novel domain may be involved in the oxidation of the initial 2-methylthiazoline ring to a 2-methylthiazole (Figure 4).

Chain elongation: KS and AT domains

After formation on EPOS P, the 2-methylthiazole ring starter unit is transferred to the PKS EPOS B, encoded by *epoB*, for subsequent elongation. EPOS B (one module), and beyond EPOS C (four modules), EPOS D (two modules) and EPOS E (one module) catalyze the incorporation of eight further carboxylic acid building blocks and their appropriate modifications to generate the olefinic spacer and the 16-membered ring of epothilones (Figure 3). Colinearity of the order of modules within the polypeptides with the order of their utilization during biosynthesis is a general feature of PKS, and reflects the processivity of these enzymes [44]. In the case of the epothilone synthase, colinearity with the biosynthetic sequence also extends to the linear order of the genes coding for the PKS, as was seen previously in other cases such as the erythromycin [45,46] and rifamycin [23,24] clusters. Colinearity was not found to extend to the level of the gene order in other examples, such as the avermectin [47], the rapamycin [48] or the FK506 clusters [49].

KS domains in the epothilone PKS exhibit 48–64% identity to type I KS domains from the databanks, and are most similar to the soraphen [16], tylosin (U78289) and pyoluteorin [50] KSs. The identities amongst the EPOS KSs are in the range 42–85%. The KSs from modules 2–8 all share the highly conserved DtaCSSsL motif (where lower case letters denote unconserved amino acids) around the active-site cysteine. This motif slightly diverges in the KS of module 1 (*Q*TACS*T*SL), and this KS also shows the lowest overall identity to the databank entries (48% maximum) and also to the other EPOS KSs (42–48%). This divergence might reflect the fact that the KS from module 1 may catalyze the condensation between the 2-methyl-4-carbonyl thiazole and a methylmalonyl group after accepting the former substrate from a PCP domain of an NRPS. All KSs, including the KSQ from module 0 (EPOS A), have the two highly conserved histidine residues [28], around 135 and 175 amino acids

Table 2

Sequence motifs of EPOS P, an NRPS.

(a)

Domain	Core	Consensus sequence	EPOS P sequence
Heterocyclization	Z1	FPL{TS}xxQxAYxxGR	FPLTdiQeSYwlGR
	Z2	RHx{IM}L(PAL)x{ND}GxQ	RHdMLRaHTlP
	C3'	(DNR)4xDxxS	DlinvDlgS
	Z3	(LI)Pxx(PAL)x{LPF}P	LPppPtLP
	Z4	(TS){PA}3x(LAF)6x(IVT)LxxW	TPtgvIlaafseYIgrW
	Z5	(GA){DQN}FT	GDFT
	Z6	P(IV)VP(TA)SxL	PVVLTSaL
	Z7	QVx{LI}Dx(QH)11xW(DYF)	QLlLDhQlyehdgdlvlaWD
Adenylation	A1	L{TS}YxEL	LTYeEL
	A2	LKAGxAY{VL}P(LI)D	LESGaAYVPID
	A3	LAYxxYTSG(ST)TGxPKG	LAYviYTSGSTGlPKG
	A4	FDxS	FDlS
	A5	NxYGPTE	SlGGATE
	A6	GELxIxGxG(VL)ARGYL	GCLyIgGvGLALGYW
	A7	Y(RK)TGDL	YKTGDL
	A8	GRxDxQVKIRGxRIELGEIE	GReDnQlKLRGyRVELGEIE
	A9	LPxYM(IV)P	LPeYMVP
	A10	NGK{VL}DR	NGKVDR
Thiolation	T	DxFFxxLGG(HD)S{LI}	QqSFvdLGATSI

(b)

Binding pocket	Position									
	235	236	239	278	299	301	322	330	331	517
Cys(2) consensus	D	L	Y	N	L	S	L	I	W	K
EPOS P	D	L	Y	N	L	S	L	I	W	K

(a) Core sequence motifs [12] in the NRPS EPOS P. Identity with the consensus sequence at the highly conserved positions within the core sequences is indicated by highlighting, sequence divergence is shown in red. Amino acids at nonconserved positions within the core motifs are shown in lower case letters. **(b)** The specificity-conferring code of the EPOS P adenylation domain. Amino acid positions refer to the PheA domain [43]. Identity with the Cys(2) consensus sequence [42] is indicated by highlighting, proposed 'wobble-like' positions within the consensus are blue.

carboxy-terminal of the active-site cysteine. The highly conserved His340 of KAS II from *E. coli*, corresponding to the last histidine in the KSs of PKS, has been implicated as a catalytic base [51].

Consensus motifs [34] and a collection of conserved amino acids in connection with overall sequence alignment [23,52] can be used to predict the broad substrate specificities of ATs. Thus, the ATs that load malonyl-CoA (or, in the case of some ATs in starter modules, acetyl-CoA), and the ATs that load longer-chain alkyl or possibly hydroxy or methoxymalonyl-CoA fall into two distinct categories. The AT domains of the epothilone PKS can also be classified into two groups using the above criteria. ATs from modules 0, 2, 3, 4 and 8 are predicted to incorporate acetate, whereas those from modules 1, 5, 6 and 7 are expected to incorporate propionate units. These assignments fit the overall structure of epothilone A.

The AT from module 3 is predicted to be specific for malonyl-CoA and hence would incorporate an acetate corresponding to C12–C11. The nascent polyketide chain of epothilones would therefore correspond to epothilone A (Figure 3). Consequently, the biosynthesis of epothilone B (with a methyl sidechain at C12) would either require a post-PKS C-methylation event, or more likely would involve a certain level of promiscuity in the selection of the substrate by this AT. Module 7 contains an AT specific for methylmalonyl-CoA; the formation of the gem-dimethyl functionality in the corresponding position of epothilones will be discussed later.

Reductive and dehydratase activities and chain termination

All of the epothilone PKS modules contain domains predicted to modify the growing polyketide chain. KR domains are present in all but module 0 (see above). The consensus motifs GxGxxGxxxA and Lx(S,G)Rx(G,T,A) and the invariant arginine that together have been implicated in forming the NADP(H)-binding site [53] are conserved in all of the epothilone KRs.

The KR of module 7 contains a Ser→Asp substitution in the Lx(*S*,G)Rx(G,T,A) motif (Lv*d*RgG). Out of 76 KR domains available for sequence comparison, this is the only example of an aspartate residue in this position. Module 3 of the pikromycin PKS contains a Ser→Thr substitution, but this module is predicted to be inactive [30]. KR 7 also has nonconservative replacements in two other highly conserved regions. About 30 amino acids carboxy-terminal to the invariant arginine, the motif Dxx(*D,E*) is replaced by Dva*r*, and the motif *V*x*H*x*A*(G,A), situated approximately 20 amino acids further carboxy-terminal is substituted by *ivyvd*G. These alterations, together with the fact that epothilones contain a keto group at C5, suggest that the KR in module 7 might be inactive, although the role, if any, of the above signature motifs in catalysis remains to be proven.

Module 3 contains an active KR, but no other reductive activities. Consequently, the β-keto group at C13 is reduced to a hydroxyl group in the nascent polyketide chain. The biosynthesis of the minor epothilone analogs C and D that feature a double bond at C13–C12 [54] would therefore require a post-PKS dehydration by a host enzyme apparently encoded outside of the sequenced region. The formation of the major epothilones A and B with the epoxide ring at C13–C12 involves a post-PKS oxidation, in which EPOS F conceivably plays a role (see below).

Modules 1, 4, 5, 7 and 8 each contain a dehydratase (DH) domain with the conserved active-site motif LxxHxxxGxxxxP [46,55]. The presence of a phenylalanine adjacent to the carboxy-terminal proline within this motif has been suggested to render the module 7 DH in rifamycin synthase inactive [35]. However, three out of the five epothilone DHs (modules 1, 5 and 8), the activity of two of which is reflected in the final structure of epothilones, contain phenylalanine at this position, as does the apparently active module 2 DH of the niddamycin cluster [29]. The activity of the apparently functional DH in module 8 of the epothilone PKS seems superfluous, because epothilones contain a hydroxyl group

at C3, although temporary dehydration at this point might be required to aid the correct folding of the polyketide chain. Discrepancies between the deduced reductive capabilities of PKS modules and the redox state of the corresponding positions in the final polyketide products are often encountered [23,24,48]. The presence of a keto group at C5 in epothilones suggests that the ostensibly functional DH of module 7 is also redundant, although this activity might play a role in the formation of the gem-dimethyl group at C4 (see the Discussion section).

The ER domains of modules 4 and 5 contain the signature motif GGVGxAAxQxA, which is believed to be involved in the formation of the NAD(P)H-binding site [53]. The same motif in the module 8 ER domain contains some very unusual amino acid substitutions (GGVGl*c*Av*r*wA) that probably render this domain inactive. Accordingly, enoyl reduction at the corresponding carbon (C3) to the alkane is not necessary for epothilone biosynthesis.

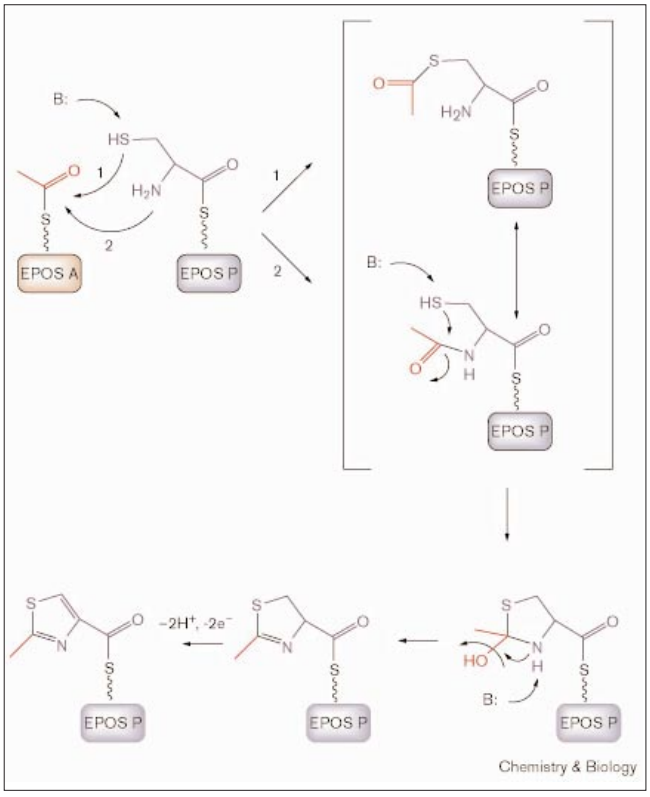
Chain termination of the epothilone PKS may require a thioesterase (TE), which releases the finished polyketide chain. A TE domain has been identified adjoining module 8 on the carboxyl terminus of EPOS E, displaying the conserved motifs GxSxD and PGdH [46].

Formation of the gem-dimethyl functionality

A striking feature of epothilones is the presence of the gem-dimethyl at C4 of the macrolide ring. Module 7 in EPOS D is predicted to incorporate a propionate unit, providing one methyl sidechain at C4. This module also contains an additional domain, integrated into the PKS between the DH and the KR domains, that displays the three degenerate consensus motifs that are present in different S-adenosylmethionine (SAM)-dependent methyltransferases [56] (Figure 5b). The highly conserved aspartate residue that is in direct contact with SAM [57] can be found 21 amino acids carboxy-terminal to motif I.

Integrated methyltransferase (MT) domains are commonly used by NRPS to N-methylate amino acids [12], but similar activities in PKS have been discovered only recently. Thus, the fumonisin polyketide synthase from *Gibberella fujikuroi* [58], the diketide and the nonaketide synthases involved in lovastatin biosynthesis in *Aspergillus terreus* [59], two PKS (PksR and PksM) from the *Bacillus subtilis* genome project (Z99113), and the PKS portions of the hybrid PKS–NRPS proteins HMWP1 (involved in yersiniabactin biosynthesis [36,60]) and NrpS (involved in swarming motility in *Proteus mirabilis* [61]) all contain MT motifs. In the cases where the product of these synthases are known, the MT domains can be implicated in C-methylation (fumonisin, lovastatin) or C-bismethylation (yersiniabactin) of their polyketide substrates. Blast searches of databanks with the MT domain of the

Figure 4



Model for the formation of the 2-methyl-4-carboxythiazole starter unit during epothilone biosynthesis. See text for details.

epothilone synthase module 7 return all of the above PKS MTs with 25–30% identities, together with the MT3 domain from the NRPS part of the yersiniabactin HMWP1, an atypical NRPS MT catalyzing C-methylation [36]. N-methyltransferases from the yersiniabactin NRPS HMWP2 [36,60] and the pyochelin PchF [37] score significantly lower and/or the identities are confined to the core motifs (Figure 5b).

Taken together, these observations lead us to propose that the MT domain in EPOS D is involved in the creation of the gem-dimethyl functionality at C4 of epothilones by catalyzing the incorporation of the second methyl side group (C21 or C22) by C-methylation.

EPOS F and other ORFs on the perimeter of the epothilone cluster

Separated from *epoE* by 115 bp with no detectable transcriptional terminator, *epoF* (Figure 2) codes for EPOS F, a deduced protein of 420 amino acids with strong sequence similarities (up to 35% identity and 48% similarity) to cytochrome P450 oxygenases. The oxygen-binding site with the motif LLxAGx(D,E), and the heme-binding pocket with the motif GxGxHx*C*xGxxLxR around the invariable cysteine residue that coordinates the heme [62]

are well conserved (IIaAGtD and GrGpHv*C*pGvsLaR). EPOS F may be involved in the formation of the epoxide ring between C12 and C13.

Two putative transport proteins can be identified bordering the epothilone synthase genes (Figure 2). The deduced protein product of *orf3* (713 amino acids) is a highly hydrophobic protein with several possible transmembrane segments. It shows about 25% identity to Na/H^+ and $drug/H^+$ antiporters of different bacteria, including D90902 from *Synechocystis* sp., and AL078635 from the vancomycin-producer *Amycolatopsis orientalis*.

The deduced protein product of *orf14* shows 35% identity and 47% similarity over its entire length (306 amino acids) to a hypothetic protein with no proposed function from *Streptomyces lividans* (AF072709), and to the human embrionic lung protein (AF006621). It is also more distantly related to cation efflux system proteins like AE000941 from *Methanobacterium thermoautotrophicum* (25% identity, 37% similarity). Both Orf 3 and Orf14 might take part in the export of epothilones from the producing cells.

The deduced protein product of *orf2* (424 amino acids) shows strong similarities to hypothetical ORFs from *Mycobacterium tuberculosis* (Z81011; 40% identity over 344 amino acids) and *Streptomyces coelicolor* (AL031107; 36% identity over 399 amino acids). Less marked similarities (up to 29% identity) were detected with the 1,4-butanediol diacrylate esterase from *Brevibacterium linens* (AB020733) and to carboxypeptidases and DD-peptidases of different bacteria. The role of Orf2, if any, in epothilone biosynthesis is not immediately obvious.

The remaining ORFs (*orf1*, *orf4–orf5*, *orf6–orf13*, and *orf15*) show no homologies to entries in the sequence databanks.

Discussion

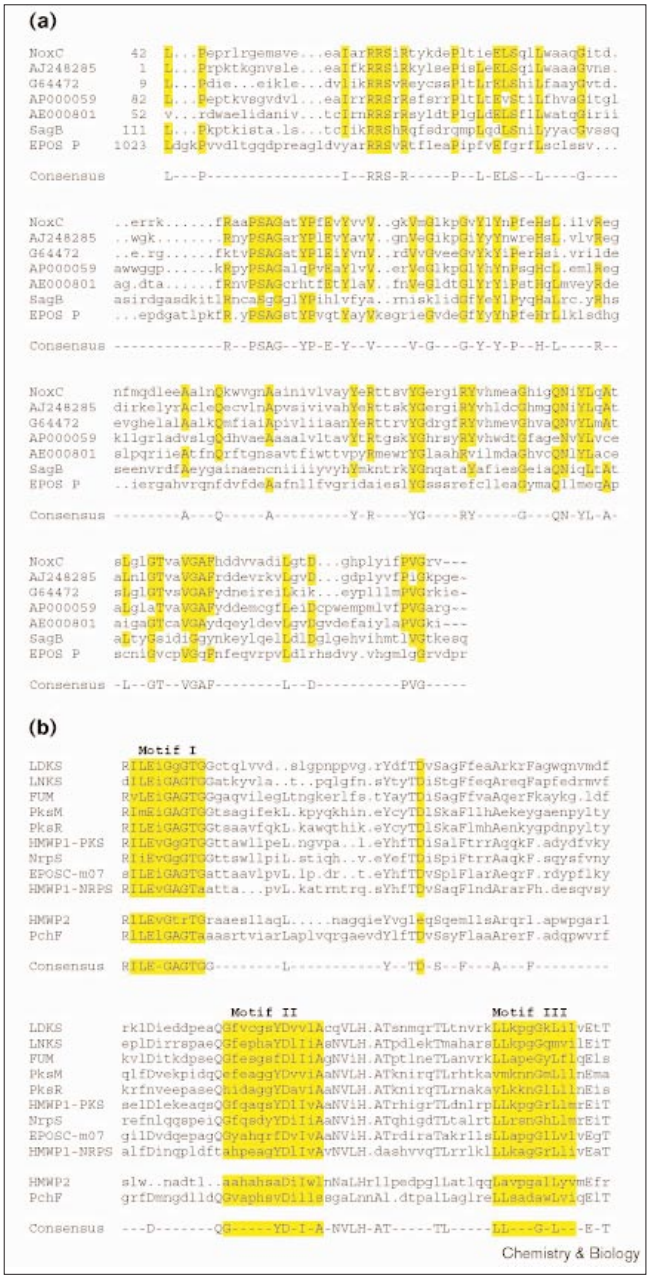
In the present work, we have identified and sequenced a contiguous region spanning 68750 bp from the genome of the epothilone producer *S. cellulosum* So ce90. This region encodes a hybrid NRPS–PKS gene cluster for the assembly of epothilones from cysteine, malonyl- and methylmalonyl-CoA, and at least one methyl group from S-adenosylmethionine (Figure 3). The deduced activities of the NRPS and PKS enzymes are in a remarkably good agreement with the final structure of epothilone A. Gene disruptions in the PKS abolished epothilone production, providing final proof for the identity of the cluster.

Probing our cosmid and Bac libraries with heterologous KS probes, and subsequent ordering of the positive clones into an overlapping clone set revealed the existence of multiple PKS clusters in the strain So ce90, which is known to produce spirangiens [10] and possibly other polyketide metabolites besides epothilone. Gene disruptions were therefore indispensable to identify the cluster responsible for epothilone production. In a recent paper, Beyer *et al.* [63] used degenerate primers to amplify PKS and NRPS sequences from myxobacteria including the epothilone producer So ce90. Although they successfully isolated cosmids that hybridized to both PKS- and NRPS-encoding amplified sequences, comparison of their published sequences with our sequence data clearly indicates that the epothilone cluster was not represented amongst their clones. The degenerate NRPS probes used in those experiments were specific to the adenylation core motif A7 and the thiolation core T [12]. Although these regions are conserved in the epothilone NRPS EPOS P, they are further separated (1350 bp in EPOS P versus 500 bp in the PCR products) by the extra domain proposed here to take part in the oxidation of the thiazoline ring to thiazole.

Based on the structure of epothilones A and B, it would appear that the epothilone PKS initiates polyketide chain assembly from a 2-methyl-4-carboxythiazole derivative. However, the sequence analysis presented here suggests that epothilone assembly begins with the loading of an acetate unit onto the ACP domain of EPOS A. The first condensation step is probably catalyzed by the heterocyclization domain of the NRPS EPOS P and involves acetyl transfer from the ACP domain of EPOS A to the PCP-bound cysteine substrate of EPOS P. The condensing nucleophile could be either the thiolate anion or the amino moiety of the cysteinyl-PCP [27] (Figure 4). Subsequent thiazoline formation also requires the activity of the EPOS P heterocyclization domain to catalyze ring closure by cyclodehydration. A similar reaction sequence has been proposed for thiazoline ring formation by the hybrid PKS–NRPS HMWP1 (a functional equivalent of the EPOS A–EPOS P pair) of the yersiniabactin synthase [36]. Oxidation of the initial thiazoline to the thiazole by the epothilone synthase probably involves a domain recruited by EPOS P. This domain is located between the adenylation core sequences A8 and A9 [12] and does not show any homologies to other NRPS sequences. However, the domain appears to be a member of a growing family (Figure 5a) of putative oxidoreductases similar to the NADH oxidase NoxC (AJ248283) from *Pyrococcus abyssi*. Interestingly, the only genetically characterized NRPS modules catalyzing thiazole formation (the bleomycin synthetase modules PTS-2 and PTS-1, [40]) apparently rely on a distinct oxidase encoded within the gene cluster to carry out this oxidation. In the absence of a published sequence, we could not check whether PTS-1 and 2 carried an oxidoreductase domain homologous to the one in EPOS P. The presence of an ER domain within EPOS A raises the alternative possibility that this domain might take part in the oxidation of the thiazoline. In the yersiniabactin system, the reduction that generates the thiazolidine from the thiazoline ring synthesized by the second NRPS module of HMWP2 was proposed to be achieved

Figure 5

(a) Alignment of the proposed oxidase domain of EPOS P with NAD(P)H oxidoreductases. Conserved amino acids (at least five identical out of seven) are highlighted. NoxC, NADH oxidase from *Pyrococcus abyssi* (AJ248283); AJ248285, hypothetical protein from *Pyrococcus abyssi*; G64472, NADH oxidase from *Methanococcus jannaschii*; AP000059, hypothetical protein from *Aeropyrum pernix*; AE000801, hypothetical protein from *Methanobacterium thermoautotrophicum*; SagB, *Streptococcus pyogenes* SagB required for streptolysin S activity; EPOS P, the putative oxidase domain of EPOS P. **(b)** Alignment of the conserved region of methytransferase (MT) domains from multifunctional proteins encompassing the SAM-dependent MT signature motifs I, II and III [56]. The signature motifs and the highly conserved aspartic acid residue (see text) are highlighted. Conserved amino acids (at least 7 identical out of 11) are indicated by capital letters. LDKS and LNKS, MTs from the lovastatin diketide and nonaketide synthases of *Aspergillus terreus* [59]; FUM, MT from the fumonisin PKS of *Gibberella fujikuroi* [58]; PksM and PksR, MTs from two PKS from the *Bacillus subtilis* genome project (Z99113); HMWP1-PKS, MT from the PKS portion of the yersiniabactin synthase HMWP1 [36,60]; NrpS, MT from the *Proteus mirabilis* NrpS involved in swarming motility [61]; EPOSC-m07, MT from module 7 of EPOS C; HMWP1-NRPS, MT from the NRPS portion of HMWP1; HMWP2, MT from the yersiniabactin NRPS HMWP2 [36,60]; PchF, MT from the pyochelin synthase PchF [37].

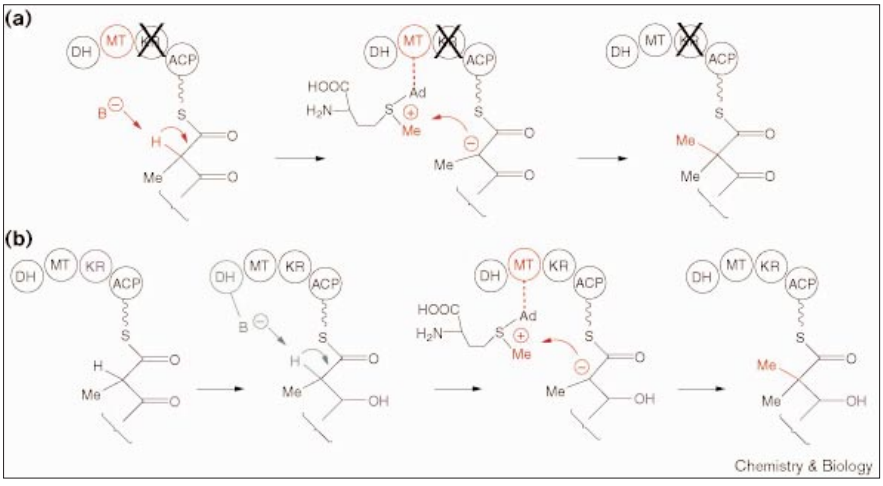


by the KR domain of the PKS module of HMWP1, acting on a C=N substrate [36].

Following thiazole formation by EPOS P, the 2-methyl-4-carbonylthiazole substrate is transferred onto the module 1 KS domain of EPOS B, and typical PKS/NRPS chain extension resumes. An increasing number of gene clusters has been described in the literature that feature a polypeptide or polyketide substrate, tethered as a thioester on the phosphopantheteine arm of a peptide carrier protein (PCP) or an ACP domain, which is transferred between NRPS and PKS platforms. Thus, the peptide chains of yersiniabactin [36] and bleomycin [40] are interrupted by a single acetate extender unit, and one of the two peptide chains of mycobactin start with a diketide [41]. Conversely, the polyketide chains of rapamycin and FK506 are terminated by condensation with pipecolate [49, 64], and the macrolide antibiotic TA from *Myxococcus xanthus* is initiated by incorporation of a glycine [17]. In the latter case, however, we would prefer to propose that a PKS-derived diketide is extended by a glycine on a NRPS before polyketide extensions resume. Gene clusters featuring hybrid NRPS–PKS systems with unknown products have also been described [61,65].

Chain extension on the EPOS PKS involves acyltransferases that can be classified into two categories on the basis of their deduced substrate specificities [34,52]. The great majority of complex polyketides synthesized by type I modular PKS of Actinomycetes are derived from both acetate and propionate (or occasionally longer chain acyl or possibly hydroxyacyl or methoxyacyl) units as demonstrated by isotopic-labeling stu dies. A few actinomycete polyketides, however, are derived exclusively from acetate extender units and then modified by S-adenosylmethionine-dependent C-methylation (see, for example, aplasmomycin or boromycin [66]). This latter, probably more ancient, route seems to be much more prevalent amongst myxobacterial complex polyketides (see, for example, myxopyronin, myxovirescin or sorangicin) [1,2]. The gene cluster for the 'variable extender'-type macrolide soraphen from *Sorangium cellulosum* So ce26 clearly coded for PKSs harboring ATs with disparate substrate specificities [16]. The first sequenced PKS module from the mixed NRPS–PKS *ta1* gene involved in the biosynthesis of the 'all-acetate'-type polyketide antibiotic TA, however, lacks an AT domain altogether [17]. It is tempting to speculate

Figure 6



Models for the biosynthesis of the gem-dimethyl moiety of epothilones. See text for details.

that type I modular PKS producing all-acetate polyketides may not need integrated AT domains and that their cognate ACP domains may self-acylate or may recruit existing ATs from other pathways, in analogy with ACPs from type II PKS [67,68].

Several natural products, such as yersiniabactin, borophycin, aplasmomycin and disorazole, have a gem-dimethyl functional group. In these compounds, the gem-dimethyl is synthesized by two successive C-methylations in which the added methyl groups originate from SAM [36,66]. Although it may have been expected that the same biosynthetic route was operative in the epothilone pathway, sequence analysis suggests a slightly different origin for the C4 gem-dimethyl functionality of epothilones. Module 7 of the epothilone synthase contains an AT specific for methylmalonyl-CoA implying that one methyl group of the gem-dimethyl is derived from the propionate extender unit. The second methyl group is believed to be introduced by the integrated C-methyltransferase domain identified within module 7.

Mechanistically, C-methylation requires a carbanion formation followed by nucleophilic attack on the methyl group of SAM. Two reaction schemes for C4 methylation can be postulated based on the identification of functional domains within module 7. If, as suggested by sequence analysis of the module 7 KR domain, the reduction of the β-carbonyl group does not occur, the α carbon could easily be deprotonated to form the reactive nucleophile susceptible to SAM-dependent methylation (Figure 6a). Alternatively, if the module 7 KR is catalytically competent, the resulting β-hydroxyacyl-S-ACP could be deprotonated by the supposedly active DH in module 7. After carbanion formation at C4, methylation instead of hydroxyl elimination would produce the gem-dimethyl group and not the expected α,β-unsaturated acyl intermediate (Figure 6b). This route, however, would result in an epothilone carbon skeleton containing a C5-hydroxyl and require a post-PKS corrective oxidation to form the observed C5-keto functionality present in epothilones.

Significance

Epothilones, potent microtubule-stabilizing agents and paclitaxel mimetics with favorable water solubility and resistance profiles, have generated great interest since their Taxol-like mode of action on microtubules was described in 1995 [4]. Although total syntheses of epothilones A and B, and a growing number of derivatives, have been achieved [9], the details of their biosynthesis received much less attention. This work describes the gene cluster responsible for the biosynthesis of epothilones in *Sorangium cellulosum* So ce90 from cysteine, malonyl- and methylmalonyl-CoA, and at least one methyl group from S-adenosylmethionine. The cluster encodes a hybrid polyketide synthase/nonribosomal peptide synthetase system harboring 48 enzymatic domains in six multifunctional polypeptides. Analysis of the cluster revealed the enzymatic logic behind the formation of the unusual 2-methylthiazole ring, polyketide chain assembly, and the formation of the gem-dimethyl functionality. The modular architecture of the synthase lends itself to modifications through combinatorial biosynthesis to create improved epothilones and provides interesting building blocks for *de novo* polyketide synthase engineering. The remarkably compact and elegantly self-contained architecture of the gene cluster also suggests that epothilones could perhaps be heterologously expressed in microorganisms more amenable to industrial production.

Materials and methods

Plasmids, bacterial strains and cultivation

Sorangium cellulosum So ce90 (DSM 6773) was grown at 30°C on agar plates of SolE medium (0.35% glucose, 0.05% tryptone, 0.15% $MgSO_4 \times 7H_2O$, 0.05% ammonium sulfate, 0.1% $CaCl_2$, 0.006% K_2HPO_4, 0.01% sodium dithionite, 0.0008% Fe-EDTA, 1.2% HEPES, 3.5% [vol/vol] supernatant of sterilized stationary *S. cellulosum* culture; pH 7.4). Liquid cultures were grown in G51t medium (0.2% glucose, 0.5% starch, 0.2% tryptone, 0.1% Probion S, 0.05% $CaCl_2 \times 2H_20$, 0.05% $MgSO_4 \times 7H_20$, 1.2% HEPES, pH 7.4) incubated at 30°C with shaking at 225 rpm. The spontaneous streptomycin-resistant mutant *So. cellulosum* BCE28/2 used for gene disruption was selected on SolE medium supplemented with 100 μg/ml streptomycin.

E. coli DH10B was used for routine cloning and Bac library construction, XL1 Blue MR for cosmid library construction, and ED8767 [69] containing the helper plasmid pUZ8 [70] as a conjugation donor.

A cosmid library of *S. cellulosum* So ce90 was raised in ScosTriplex-II [25], and the Bac library was made in pBelobacII [22]. pCIB132 [16] was used as the mobilizable plasmid in conjugation experiments, and pBluescript II SK for routine cloning. pCIB132 derivatives used for the insertional inactivation of the epothilone PKS (Figure 2) carried the following fragments derived from the Bac clone b15: *Bam*HI fragment 1 (sequence position 16876–20779), *Bam*HI fragment 2 (20779–22986), *Bam*HI fragment 3 (42528–44755), *Cla*I fragment 4 (7692–10464).

Genetic procedures

Standard genetic techniques for *E. coli* and for in vitro DNA manipulation were as described [71]. Isolation of total DNA from *So. cellulosum*, matings between *E. coli* and *S. cellulosum* for plasmid transfer, and selection of transconjugants followed established protocols [16,72]. BCE28/2, a streptomycin resistant spontaneous mutant of *So. cellulosum* So ce90 was used for the conjugations to select with streptomycin against the *E. coli* partners after mating. To generate a Bac library, *S. cellulosum* cells were embedded into agarose blocks, lysed, and the liberated genomic DNA was partially digested with the restriction enzyme *Hin*dIII. The digested DNA was separated on an agarose gel by pulsed-field electrophoresis. Approximately 90–150 kb DNA fragments were isolated and ligated into the vector pBelobacII. The ligation mixture was used to transform *E. coli* DH10B electrocompetent cells using standard electroporation techniques. Chloramphenicol-resistant recombinant colonies were transferred to a positively charged nylon membrane filter in 384 3×3 grid format. The clones were lysed and the DNA was cross-linked to the filters. The same clones were also preserved as liquid cultures at −80°C. For the cosmid library of *S. cellulosum* So ce90, genomic DNA was partially digested with *Sau*3AI, and size-selected fragments (approximately 40 kb) were ligated into ScosTriplex II digested with *Bam*HI and *Xba*I. The ligation mixture was packaged with Gigapack III XL (Stratagene) and used to transfect *E. coli* XL1 Blue MR cells. Digoxigenin-labeled probes were prepared and colony and Southern hybridizations were done as recommended by the manufacturer of the DIG system (Boehringer Mannheim).

DNA sequencing and analysis

Overlapping clone sets were generated by analyzing complete *Bam*HI and *Sac*I digests of Bacs and cosmids using the programs FPC [73] and IMAGE [74]. Automated DNA sequencing was done on double-stranded DNA templates by the dideoxynucleotide chain termination method, using Applied Biosystems model 377 sequencers. Subclones used in the initial gene disruptions were sequenced by primer walking, and the sequence was compiled with the program SEQUENCHER (Gene Codes Corporation). Inserts from subclones of the epothilone cluster (Figure 2) were subjected to random fragmentation using a Hydroshear apparatus (Genomic Instrumentation Services, Inc.) to yield an average fragment size of 1–2 kb. The fragments were end-repaired using T4 DNA Polymerase and Klenow DNA Polymerase in the presence of deoxynucleotide triphosphates and phosphorylated with T4 DNA Kinase. Alternatively, inserts from subclones were partially digested with *Hae*III to yield fragments with an average size of 1–2 kb. Fragments in the size range of 1.5–2.2 kb were isolated from agarose gels and ligated into pBluescript II SK that had been cut with *Eco*RV and dephosphorylated. Random subclones were sequenced using the universal forward and reverse sequencing primers. The chromatograms were analyzed and the nucleotide sequences were assembled with the PHRED, PHRAP and CONSED programs [75–77]. To fill sequence gaps, resolve discrepancies, clarify low-quality regions, and cover clone junctions, custom-designed oligonucleotide primers were used for sequencing on either the original cosmids or selected clones from the random subclone libraries. Both strands were completely sequenced, and every basepair is covered with an aggregated PHRED score of at least 40 (confidence level of 99.99%). The University of Wisconsin Genetics Computer Group programs [78] were used for sequence analysis.

Assay for epothilone production

Transconjugant colonies of *S. cellulosum* were streaked out and cultivated on S42 agar [72] containing streptomycin (30 mg/l), phleomycin (30 mg/l) and kanamycin (50 mg/l) for a second round of selection and purification. After incubation at 30°C for 7 days cells grown on about 1 cm^2 surface of the selective plates were transferred into 10 ml of G52-H medium (0.2% glucose, 0.8% potato starch, 0.2% defatted soya meal, 0.2% yeast extract (low salt), 0.1% $CaCl_2 \times 2\,H_2O$, 0.1 % $MgSO_4 \times 7\ H_2O$, 0.0008 % Fe-EDTA sodium salt, pH ad 7.4 with NaOH) in a 50 ml Erlenmeyer flask. After incubation at 30°C and 180 rpm for 7 days, the culture was transferred into 50 ml of medium G52-H in a 200 ml Erlenmeyer flask. After incubation at 30°C and 180 rpm for 3–4 days, 10 ml of this culture was transferred into 50 ml of medium 23B3 (0.2% glucose, 2% potato starch, 1.6% defatted soya meal, 0.0008% Fe-EDTA Sodium salt, 0.5% HEPES (4-(2-hydroxyethyl)-piperazine-1-ethane-sulfonic-acid), 2% vol/vol polysterole resin XAD16 (Rohm & Haas), pH ad 7.8 with KOH) in a 200 ml Erlenmeyer flask. Quantitative determination of the epothilones produced takes place after incubation of the cultures at 30°C and 180 rpm for 7 days. The resin was harvested from the culture on a nylon sieve, resuspended in 10 ml isopropanol and extracted by shaking at 180 rpm for 1 h. 1 ml is removed from this suspension and centrifuged at 12,000 rpm. Epothilones A and B were detected in the supernatant with a Hewlett Packard HPLC with a diode array ultraviolet detector set at 250 nm, fitted with a Waters-Symetry C18 column (3.5 μm; 4.6×100 mm). Elution was with a step gradient (1 ml/min.) of 0.02% phosphoric acid and acetonitrile (acetonitrile concentration: 30% 0–10 min., 60% 10–11 min., 100%, 11–14 min., 30% 14–18 min.). Retention times for epothilones A and B were 8.5–9 min. and 9.5–10 min, respectively.

Accession number

The entire sequence reported here have been deposited in the GenBank database under the accession number AF210843.

Acknowledgements

The authors are indebted to Sharon Potter for her help in computer analyses and thank K.-H. Altmann and F. Petersen for supporting this work.

References

1. Reichenbach, H. & Höfle, G. (1999). Myxobacteria as producers of secondary metabolites. In *Drug Discovery from Nature* (Grabley, S. & Thiericke+, R., eds), pp.149-179, Springer-Verlag, Berlin Heidelberg.
2. Reichenbach, H. & Höfle, G. (1993). Production of bioactive secondary metabolites. In *Myxobacteria II* (Dworkin, M. and Kaiser, D., eds), pp.347-397, American Society for Microbiology, Washington, DC.
3. Gerth, K., Bedorf, N., Höfle, G., Irschik, H. & Reichenbach, H. (1996). Epothilones A and B: antifungal and cytotoxic compounds from *Sorangium cellulosum* (Myxobacteria). *J. Antibiot.* **49**, 560-563.
4. Bollag, D.M., *et al.*, & Woods, C.M. (1995). Epothilones, a new class of microtubule-stabilizing agents with a taxol-like mechanism of action. *Cancer Res.* **55**, 2325-2333.
5. Gunasekera, S.P., Gunasekera, M., Longley, R.E. & Schulte, G.K. (1990). Discodermolide: a new bioactive polyhydroxylated lactone from the marine sponge *Discodermia dissoluta*. *J. Org. Chem.* **55**, 4912-4915.
6. Hung, D.T., Chen, J. & Schreiber, S.L. (1996). (+)-Discodermolide binds to microtubules in stoichiometric ratio to tubulin dimers, blocks taxol binding and results in mitotic arrest. *Chem. Biol.* **3**, 287-293.

7. Su, D.-S., *et al.*, & Horwitz, S.B. (1997). Structure-activity relationships of the epothilones and the first *in vivo* comparison with paclitaxel. *Angew. Chem. Int. Ed. Engl.* **36**, 2093-2096.
8. Kowalski, R.J., Giannakakou, P. & Hamel, E. (1997). Activities of the microtubule-stabilizing agents epothilones A and B with purified tubulin and in cells resistant to paclitaxel (Taxol®). *J. Biol. Chem.* **272**, 2534-2541.
9. Nicolaou, K.C., Roschangar, F. & Vourloumis, D. (1998). Chemical biology of epothilones. *Angew. Chem. Int. Ed. Engl.* **37**, 2014-2045.
10. Höfle, G., *et al.*, & Reichenbach, H. (1996). Antibiotics from gliding bacteria. 77. Epothilone A and B - novel 16-membered macrolides with cytotoxic activity: isolation, crystal structure, and conformation in solution. *Angew. Chem. Int. Ed. Engl.* **35**, 1567-1569.
11. Cane, D.E., Walsh, C.T. & Khosla, C. (1998). Harnessing the biosynthetic code: Combinations, permutations, and mutations. *Science* **282**, 63-68.
12. Konz, D. & Marahiel, M.A. (1999). How do peptide synthetases generate structural diversity? *Chem. Biol.* **6**, R39-R48.
13. Hopwood, D.A. (1997). Genetic contributions to understanding polyketide synthases. *Chem. Rev.* **97**, 2465-2497.
14. Marahiel, M.A., Stachelhaus, T. & Mootz, H.D. (1997). Modular peptide synthetases involved in non-ribosomal peptide synthesis. *Chem. Rev.* **97**, 2651-2673.
15. von Döhren, H., Keller, U., Vater, J. & Zocher, R. (1997). Multifunctional peptide synthetases. *Chem. Rev.* **97**, 2675-2705.
16. Schupp, T., *et al.*, & Ligon, J.M. (1995). A *Sorangium cellulosum* (Myxobacterium) gene cluster for the biosynthesis of the macrolide antibiotic soraphen A: cloning, characterization, and homology to polyketide synthase genes from actinomycetes. *J. Bacteriol.* **177**, 3673-3679.
17. Paitan, Y., Alon, G., Orr, E., Ron, E.Z. & Rosenberg, E. (1999). The first gene in the biosynthesis of the polyketide antibiotic TA of *Myxococcus xanthus* codes for a unique PKS module coupled to a peptide synthetase. *J. Mol. Biol.* **286**, 465-474.
18. Pospiech, A., Bietenhader, J. & Schupp, T. (1996). Two multifunctional peptide synthetases and an *O*-methyltransferase are involved in the biosynthesis of the DNA-binding antibiotic and antitumour agent saframycin Mx1 from *Myxococcus xanthus*. *Microbiology* **142**, 741-746.
19. Hutchinson, C.R. (1998). Combinatorial biosynthesis for new drug discovery. *Curr. Opin. Microbiol.* **1**, 319-329.
20. Carreras, C.W. & Santi, D.V. (1998). Engineering of modular polyketide synthases to produce novel polyketides. *Curr. Opin. Biotechnol.* **9**, 403-411.
21. Staunton, J. (1998). Combinatorial biosynthesis of erythromycin and complex polyketides. *Curr. Opin. Chem. Biol.* **2**, 339-345.
22. Wang, K., Boysen, C., Shizuya, H., Simon, M.I. & Hood, L. (1997). Complete nucleotide sequence of two generations of a bacterial artificial chromosome cloning vector. *BioTechniques* **23**, 992-994.
23. August, P.R., *et al.*, & Floss, H.G. (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.
24. Schupp, T., Toupet, C., Engel, N. & Goff, S. (1998). Cloning and sequence analysis of the putative rifamycin polyketide synthase gene cluster from *Amycolatopsis mediterranei*. *FEMS Microbiol. Lett.* **159**, 201-207.
25. Ji, H., Francisco, T., Smith, L.M. & Guilfoyle, R.A. (1996). Rapid restriction mapping of cosmids by sequence-specific triple-helix-mediated affinity capture. *Genomics* **31**, 185-192.
26. Wright, F. & Bibb, M.J. (1992). Codon usage in the G + C-rich Streptomyces genome. *Gene* **113**, 55-65.
27. Roy, S.R., Gehring, A.M., Milne, J.C., Belshaw, P.J. & Walsh, C.T. (1999). Thiazole and oxazole peptides: biosynthesis and molecular machinery. *Nat. Prod. Rep.* **16**, 249-263.
28. Aparicio, J.F., *et al.*, & Leadlay, P.F. (1996). Organization of the biosynthetic gene cluster for rapamycin in *Streptomyces hygroscopicus*: analysis of the enzymatic domains in the modular polyketide synthase. *Gene* **169**, 9-16.
29. Kakavas, S.J., Katz, L. & Stassi, D. (1997). Identification and characterization of the niddamycin polyketide synthase genes from *Streptomyces caelestis*. *J. Bacteriol.* **179**, 7515-7522.
30. Xue, Y., Zhao, L., Liu, H.-W. & Sherman, D.H. (1998). A gene cluster for macrolide antibiotic biosynthesis in *Streptomyces venezuelae*: Architecture of metabolic diversity. *Proc. Natl Acad. Sci. USA* **95**, 12111-12116.
31. Kuhstoss, S., Huber, M., Turner, J.R., Paschal, J.W. & Rao, R.N. (1996). Production of a novel polyketide through the construction of a hybrid polyketide synthase. *Gene* **183**, 231-236.
32. Bisang, C., *et al.*, & Leadlay, P.F. (1999). A chain initiation factor common to both modular and aromatic polyketide synthases. *Nature* **401**, 502-505.
33. Aparicio, J.F., Colina, A.J., Ceballos, E. & Martín, J.F. (1999). The biosynthetic gene cluster for the 26-membered ring polyene macrolide pimaricin. *J. Biol. Chem.* **274**, 10133-10139.
34. Haydock, S., *et al.*, & Leadlay, P.F. (1995). Divergent structural motifs correlated with the substrate specificity of (methyl)malonyl-CoA: acyl carrier protein transacylase domains in modular polyketide synthase. *FEBS Lett.* **374**, 246-248.
35. Tang, L., Yoon, Y.J., Choi, C.-Y. & Hutchinson, C.R. (1998). Characterization of the enzymatic domains in the modular polyketide synthase involved in rifamycin B biosynthesis by *Amycolatopsis mediterranei*. *Gene* **216**, 255-265.
36. Gehring, A.M., *et al.*, & Perry, R.D. (1998). Iron acquisition in plague: modular logic in enzymic biogenesis of yersiniabactin by *Yersinia pestis*. *Chem. Biol.* **5**, 573-586.
37. Reimmann, C., Serino, L., Beyeler, M. & Haas, D. (1998). Dihydroaeruginoic acid synthetase and pyochelin synthetase, products of the *pch*EF genes, are induced by extracellular pyochelin in *Pseudomonas aeruginosa*. *Microbiology* **144**, 3135-3148.
38. Konz, D., Klens, A., Schorgendorfer, K. & Marahiel, M.A. (1997). The bacitracin biosynthesis operon of *Bacillus licheniformis* ATCC 10716: molecular characterization of three multi-modular peptide synthetases. *Chem. Biol.* **4**, 927-937.
39. Tolmasky, M.E., Actis, L.A. & Crosa, J.H. (1993). A single amino acid change in AngR, a protein encoded by pJM1-like virulence plasmids, results in hyperproduction of anguibactin. *Infect. Immun.* **61**, 3228-3233.
40. Shen, B., Du, L., Sanchez, C., Chen, M. & Edwards, D.J. (1999). Bleomycin biosynthesis in *Streptomyces verticillus* ATCC15003: a model of hybrid peptide and polyketide biosynthesis. *Bioorg. Chem.* **27**, 155-171.
41. Quadri, L.E.N., Sello, J., Keating, T.A., Weinreb, P.H. & Walsh, C.T. (1999). Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem. Biol.* **5**, 631-645.
42. Stachelhaus, T., Mootz, H.D. & Marahiel, M.A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493-505.
43. Conti, E., Stachelhaus, T., Marahiel, M.A. & Brick, P. (1997). Structural basis for the activation of phenylalanine in the nonribosomal biosynthesis of gramicidin S. *EMBO J.* **16**, 4174-4183.
44. Yu, T.-W., *et al.*, & Floss, H.G. (1999). Direct evidence that the rifamycin polyketide synthase assembles polyketide chains processively. *Proc. Natl Acad. Sci. USA* **96**, 9051-9056.
45. Cortés, J., Haydock, S.F., Roberts, G.A., Bevitt, D.J. & Leadlay, P.F. (1990). An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea*. *Nature* **348**, 176-178.
46. 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.
47. Ikeda, H., Nonomiya, T., Usami, M., Ohta, T. & Omura, S. (1999). Organization of the biosynthetic gene cluster for the polyketide anthelmintic avermectin in *Streptomyces avermitilis*. *Proc. Natl Acad. Sci. USA* **96**, 9509-9514.
48. Schwecke, T., *et al.*, & Leadlay, P.F. (1995). The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proc. Natl Acad. Sci. USA* **92**, 7839-7843.
49. Motamedi, H. & Shafiee, A. (1998). The biosynthetic gene cluster for the macrolactone ring of the immunosuppressant FK506. *Eur. J. Biochem.* **256**, 528-534.
50. Nowak-Thompson, B., Chaney, N., Wing, J.S., Gould, S.J. & Loper, J.E. (1999). Characterization of the pyoluteorin biosynthetic gene cluster of *Pseudomonas fluorescens* Pf-5. *J. Bacteriol.* **181**, 2166-2174.
51. Huang, W., *et al.*, & Lindquist, Y. (1998). Crystal structure of β-ketoacyl-acyl carrier protein synthase II from *E. coli* reveals the molecular architecture of condensing enzyme. *EMBO J.* **17**, 1183-1191.
52. Katz, L. (1997). Manipulation of modular polyketide synthases. *Chem. Rev.* **97**, 2557-2575.
53. Scrutton, N.S., Berry, A. & Perham, R.N. (1990). Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. *Nature* **343**, 38-43.

54. Reichenbach, H., Höfle, G., Gerth, K. & Steinmetz, H. (1998). Epothilone C, D, E and F, production process, and their use as cytostatics well as phytosanitary agents. *PCT Int. Appl.* WO 9822461.
55. Bevitt, D.J., Cortés, J., Haydock, S.F. & Leadlay, P.F. (1992). 6-Deoxyerythronolide B synthase 2 from *Saccharopolyspora erythraea*. Cloning of the structural gene, sequence analysis and inferred domain structure of the multifunctional enzyme. *Eur. J. Biochem.* **204**, 39-49.
56. Kagan, R.M. & Clarke, S. (1994). Widespread occurrence of three sequence motifs in diverse S-adenosylmethionine-dependent methyltransferases suggests a common structure for these enzymes. *Arch. Biochem. Biophys.* **310**, 417-427.
57. Schucklebier, G., O'Gara, M., Sänger, W. & Cheng, X.D. (1995). Universal catalytic domain structure of AdoMet-dependent methyltransferases. *J. Mol. Biol.* **247**, 16-20.
58. Proctor, R.H., Desjardins, A.E., Plattner, R.D. & Hohn, T.M. (1999). A polyketide synthase gene required for biosynthesis of fumonisin mycotoxins in Gibberella fujikuroi mating population A. *Fungal Genet. Biol.* **1**, 100-112.
59. Kennedy, J.*, et al.*, & Hutchinson, C.R. (1999). Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. *Science* **284**, 1368-1372.
60. Pelludat, C., Rakin, A., Jacobi, C.A., Schubert, S. & Heesemann, J. (1998). The yersiniabactin biosynthetic gene cluster of *Yersinia enterolitica*: Organization and siderophore-dependent regulation. *J. Bacteriol.* **180**, 538-546.
61. Gaisser, S. & Hughes, C. (1996). A locus coding for putative non-ribosomal peptide/polyketide synthase functions is mutated in a swarm-defective *Proteus mirabilis* strain. *Mol. Gen. Genet.* **253**, 415-427.
62. Cupp-Vickery, J.R. & Poulos, T.L. (1995). Structure of the cytochrome $P450_{ERYF}$ involved in erythromycin biosynthesis. *Nature Struct. Biol.* **2**, 144-153.
63. Beyer, S., Kunze, B., Silakowski, B. & Müller, 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 epothilone producing strain *Sorangium cellulosum* So ce90. *Biochim. Biophys. Acta* **1445**, 185-195.
64. König, A.*, et al.*, & Leadlay, P.F. (1997). The pipecolate incorporating enzyme for the biosynthesis of the immunosuppressant rapamycin. *Eur. J. Biochem.* **247**, 526-534.
65. Albertini, A.M., Caramori, T., Scoffone, F., Scotti, C. & Galizzi, A. (1995). Sequence around the 159° region of the *Bacillus subtilis* genome: the *pksX* locus spans 33.6 kb. *Microbiology* **141**, 299-309.
66. Rawlings, B.J. (1997). Biosynthesis of polyketides. *Nat. Prod. Rep.* **14**, 523-556.
67. Hitchman, T.S., Crosby, J., Byrom, K.J., Cox, R.J. & Simpson, T.J. (1998). Catalytic self-acylation of type II polyketide synthase acyl carrier proteins. *Chem. Biol.* **5**, 35-47.
68. Revill, W.P., Bibb, M.J. & Hopwood, D.A. (1995). Purification of a malonyl transferase from *Streptomyces coelicolor* A3(2) and analysis of its genetic determinant. *J. Bacteriol.* **177**, 3946-3952.
69. Murray, N.E., Brammar, W.J. & Murray, K. (1977). Lambdoid phages that simplify the recovery of *in vitro* recombinants. *Mol. Gen. Genet.* **150**, 53-61.
70. Hedges, R.W. & Matthew, M. (1979). Acquisition by *Escherichia coli* of plasmid-borne β-lactamases normally confined to *Pseudomonas* spp. *Plasmid* **2**, 269-278.
71. Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989). *Molecular cloning: A laboratory manual*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
72. Jaoua, S., Neff, S. & Schupp, T. (1992). Transfer of mobilizable plasmids to *Sorangium cellulosum* and evidence for their integration into the chromosome. *Plasmid* **28**, 157-165.
73. Soderlund, C., Longden, I. & Mott, R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523-535.
74. Sulston, J., Mallett, F., Durbin, R. & Horsnell, T. (1989). Image analysis of restriction enzyme fingerprint autoradiograms. *Comput. Appl. Biosci.* **5**, 101-106.
75. Ewing, B., Hillier, L.D., Wendl, M.C. & Green, P. (1998). Base-calling automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175-183.
76. Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186-194.
77. Gordon, D., Abajian, C. & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195-202.
78. Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence-analysis programs for the VAX. *Nucleic Acids Res.* **12**, 387-395.

Because *Chemistry & Biology* operates a 'Continuous Publication System' for Research Papers, this paper has been published via the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/cmb – for further information, see the explanation on the contents pages.