ORIGINAL PAPER

A novel docking domain interface model predicting recombination between homoeologous modular biosynthetic gene clusters

Antonio Starcevic · Janko Diminic · Jurica Zucko · Mouhsine Elbekali · Tobias Schlosser · Mohamed Lisfi · Ana Vukelic · Paul F. Long · Daslav Hranueli · John Cullum

Received: 18 September 2010 / Accepted: 10 November 2010 / Published online: 24 November 2010

Abstract An in silico model for homoeologous recombination between gene clusters encoding modular polyketide synthases (PKS) or non-ribosomal peptide synthetases (NRPS) was developed. This model was used to analyze recombination between 12 PKS clusters from *Streptomyces* species and related genera to predict if new clusters might give rise to new products. In many cases, there were only a limited number of recombination sites (about 13 per cluster pair), suggesting that recombination may pose constraints on the evolution of PKS clusters. Most recombination events occurred between pairs of ketosynthase (KS) domains, allowing the biosynthetic outcome of the recombinant modules to be predicted. About 30% of recombinants were predicted to produce polyketides. Four NRPS clusters from *Streptomyces* strains were also used for in silico recombination. They yielded a comparable number of recombinants to PKS clusters, but the adenylation (A) domains contained the largest proportion of recombination events; this might be a mechanism for producing new substrate specificities. The extreme G + C-content, the presence of linear chromosomes and plasmids, as well as the lack of a *mutSL*-mismatch repair system should favor production of recombinants in *Streptomyces* species.

Keywords Polyketide synthase · Non-ribosomal peptide synthetase · *Streptomyces* · *Bacillus* · Chi sequence

Introduction

The synthesis of microbial secondary metabolites is often encoded by large gene clusters and they are probably the best system available for studying the evolution of collectives of microbial genes [16]. *Streptomyces* and related genera are prolific producers of secondary metabolites, which include many commercially important compounds: e.g., erythromycin (antibiotic), avermectin (antiparasitic), doxorubicin (anticancer), and rapamycin (immuno-suppressant). This raises the question of whether there are special features of *Streptomyces* species that promote the generation of chemical diversity. Certainly, their genome sequences encode a large number of secondary metabolite clusters and the vast chemical diversity of their products could be explained by frequency-dependent selection. Modular biosynthetic clusters, e.g., polyketide synthases (PKS) or non-ribosomal peptide synthetases (NRPS), are particularly interesting for evolutionary studies as they function according to an "assembly line" principle in

Electronic supplementary material The online version of this article (doi:10.1007/s10295-010-0909-0) contains supplementary material, which is available to authorized users.

A. Starcevic · J. Zucko · M. Elbekali · T. Schlosser · M. Lisfi · J. Cullum (✉)
LB Genetik, University of Kaiserslautern, Postfach 3049,
67653 Kaiserslautern, Germany
e-mail: cullum@rhrk.uni-kl.de

A. Starcevic · J. Diminic · J. Zucko · D. Hranueli
Section for Bioinformatics, Faculty of Food Technology
and Biotechnology, University of Zagreb,
Pierottijeva 6, 10000 Zagreb, Croatia

A. Vukelic
Section for Mathematics, Faculty of Food Technology
and Biotechnology, University of Zagreb, Pierottijeva 6,
10000 Zagreb, Croatia

P. F. Long
Pharmaceutical Science Institute, King's College London,
Franklin-Wilkins Building, Stamford Street,
London SE1 9NH, UK

which each module is usually responsible for a single extension step during the synthesis of the product and the growing chain is passed from module to module until complete. The chemistry is understood in considerable details [13, 15, 19, 31, 36] and is summarized briefly below.

For PKSs, each module contains an acyl carrier protein (ACP) domain, which covalently binds the growing polyketide chain as a thiol ester. To extend the chain, a new extender unit is transferred from coenzyme A to the ACP domain of the next module, a reaction catalyzed by the acyl transferase (AT) domain, which determines the substrate used. AT domains show different substrate specificities: malonate and methylmalonate are most common, with ethylmalonate and methoxymalonate also being known substrates. The β-ketoacyl synthase (KS) domain then transfers the growing chain to the new extender unit in a decarboxylative condensation reaction. Thus, the growing chain is passed from module to module and the backbone is extended by two carbon atoms at each step; if a substrate other than malonate is used, there will also be a side chain (methyl, ethyl, methoxy for the substrates above). These reactions would result in a polyketide with a keto group on every second carbon atom in the backbone. However, many modules carry reduction domains, resulting in a more varied chemistry. A ketoreductase (KR) domain reduces the keto group to a hydroxyl group. If a dehydratase (DH) domain is also present, a further reduction to a double bond occurs. If, in addition, an enoylreductase domain (ER) is present, there is a full reduction to an alkyl group. Further chemical diversity is introduced by the presence of chiral centers, whose stereochemistry is also controlled by the modules involved. Synthesis is initiated by transferring a starter unit from coenzyme A to the ACP domain of the "loading domain" (the starter module). Many PKSs use acetate or propionate as a starter, but others use less usual starters such as isobutyrate (avermectin) or 3-amino-5-hydroxybenzoate (rifamycin). On completion of the last extension step, the polyketide chain must be released from the synthase, often using a thioesterase (TE) domain in the last module. This usually results in cyclization of the product. The order of domains in a module is nearly invariant: KS–AT–DH–ER–KR–ACP. Modular PKSs usually consist of several polypeptides, each containing multiple modules. The modules in a particular polypeptide are used in order from the NH_2-end to the COOH-end of the protein. The ends of the polypeptides carry docking domains and specific interactions between pairs of docking domains assemble the polypeptides in the correct synthetic order.

Although they have a very different chemistry, forming C–N bonds rather than C–C bonds, modular NRPSs have a very similar organization to PKSs, allowing them to be analyzed with common computer programs. Like PKSs, each module binds the growing chain as a thiol ester attached to a peptidyl carrier protein domain (PCP). The next amino acid extender unit is loaded onto the PCP domain of the next module by an adenylation domain (A): this reaction involves an activated intermediate coupled to AMP. The growing chain is then coupled with the new extender unit by a condensation domain (C). The order of domains in a module is C–A–PCP. As for PKSs, it is possible for modules to contain additional optional domains such as epimerization domains (E) or methyl transferase domains (MT).

The sequences of the different modules are similar so that evolution by duplication or deletion of modules, or by recombination between clusters probably all contribute to the evolution of modular clusters [21]. Conjugation systems are common in *Streptomyces* strains [18] and there is evidence for horizontal gene transfer for some secondary metabolite clusters [14, 25]. Some clusters are carried on plasmids [26] and it has been shown that a cluster can be transferred from the chromosome to a plasmid [12, 17, 29]. Thus, it is likely that modular clusters derived from different *Streptomyces* strains would be offered opportunities to recombine.

Most bacteria have circular chromosomes, but *Streptomyces* species have linear chromosomes [10] as well as many linear plasmids [24]. Recombination between two circular DNA molecules usually requires a double cross-over event to generate viable recombinants, whereas a single cross over can generate viable recombinants between two linear molecules. This has been observed in *Streptomyces* species where single cross overs between the chromosome and a linear plasmid produce linear recombinant molecules with one chromosome end and one plasmid end [29, 37]. A double cross over between two non-identical sequences would probably be exceedingly rare and would replace an internal part of each cluster with sequences from another cluster. In contrast, a single cross over should be much more frequent, and will generate recombinant clusters with one end derived from each parental cluster. This process would produce radically new clusters, whereas double cross overs between two clusters, or recombination within a cluster, would produce smaller changes.

In order to evaluate the role of such single cross overs in the evolution of modular biosynthetic clusters, it is important to know what the distribution of potential recombination sites is and whether recombinants are likely to synthesize new products. Little is known about the details of homologous recombination systems in *Streptomyces* species, but it appears that homologous recombination is similar in its properties in many species. Such recombination requires a region of similar sequence in the two parental molecules and there appears to be a requirement for a minimal length of perfect identity, a minimal effective pairing sequence (MEPS); for *E. coli* the length of MEPS was measured as 23–27 bp [32]. There also appears to be a requirement for a region around the MEPS of high sequence similarity: for yeast a similarity of 74% was shown to function [11]. In *Escherichia coli*, *Bacillus subtilis*, and *Lactococcus lactis*, species whose recombination functions have been investigated in some detail, the chromosomes contain specific Chi sequences, which stimulate recombination. These sequences interact with an "ExoV"-like enzyme, which combines a DNA helicase activity and an endonuclease activity [4].

In this paper, we describe the construction of an in silico model for homoeologous recombination and its application to generate single cross overs between modular biosynthetic gene clusters. The model was used to examine the following questions:

1. Does recombination occur between most module pairs of two clusters, or are there only a limited number of possible sites? Constraints might limit the chemical diversity that can be generated.
2. Do the sites occur in domains where recombination would generate novel module specificities, or would the recombinant modules have identical specificities as one of the parent modules?
3. Do the recombinant clusters have architectures compatible with synthesis of a product, which would be subjected to selection immediately?
4. Are there differences between PKS and NRPS clusters in recombination behavior?
5. Are there differences in recombination between clusters from *Streptomyces* with high G + C-content and *Bacillus* with less extreme G + C-content?

Here we present data that provide insights into these five questions.

Materials and methods

DNA sequences (Table 1) and protein sequences were obtained from the GenBank database at NCBI (http://www.ncbi.nlm.nih.gov/). The following protein sequences were used: UvrD (BAA00048), RecB (AAC75859), PcrA (CAA75552), and AddB (CAA74481). BLAST searches [1] were carried out at NCBI.

The recombination algorithm was implemented in Perl using the EMBOSS needle program (http://www.ebi.ac.uk/Tools/emboss/align/index.html) for Needleman-Wunsch alignment [28]. It was integrated into the *ClustScan* package [34, 38] with an extension of the Java user interface (manuscript in preparation).

A program was written in Perl (http://www.perl.org/) using the BioPerl packages (http://www.bioperl.org/) to count the number of every oligomer of a given length in each strand in a region of a sequence. The length of the oligomers and the coordinates of the segment of the sequence to be scanned were input by the user. A cut-off giving the minimum number of occurrences of the oligomer to be considered (usually 100 copies) was used to eliminate rare sequences, which might show a strand bias by chance. The program outputs the oligomers with the highest strand bias. This program was used with the genomes of *Streptomyces coelicolor* A3(2) and *Streptomyces avermitilis* taking regions encompassing each chromosome arm from the telomere to the replication origin assuming an *oriC* coordinate of 4,270,000 (for *S. coelicolor*) and 5,288,500 (for *S. avermitilis*). A program was also written to count the number of a given potential Chi sequence in each strand in segments of 100 kb along the chromosome.

Results

A recombination model

The recombination model requires two properties of a pair of recombination sites:

1. There is a region of identity in the two sequences in which the recombination takes place: an Effective Pairing Sequence (EPS). The model assumes that there is a minimum length, the MEPS, but no sequence specificity is assumed.
2. The region around the EPS must have a high degree of similarity in the two parental sequences. The model assumes that a symmetric region around the EPS must show a base identity above a certain level. The model assumes that the similarity is calculated after alignment of the two regions using the Needleman-Wunsch algorithm [28], which gives a rigorous alignment allowing gaps.

The stringency of the recombination model is, thus, controlled by three parameters: the length of the MEPS, the length of the region of high similarity and the degree of identity required. A program was written to search pairs of DNA sequences for recombination sites. For standard use, the default parameters were a MEPS of 27 bp, with a 75% sequence identity in a region of ±100 bp around the EPS. The recombination algorithm was implemented as a module *CompGen* (to be described elsewhere) of the modular cluster analysis program *ClustScan* ([34, 38]; generating the new version of *ClustScan*: *ClustScan*-Professional; http://bioserv.pbf.hr/cms/), which had the advantage of using the program's functions for relating DNA

Table 1 DNA sequences used in the analyses

Sequence	Organism	GenBank accession	Length (bp)	Number of modules
Genome	*S. coelicolor*	AL645882	8,667,507	–
Genome	*S. avermitilis*	BA000030	9,025,608	–
PKS				
Amphotericin	*S. nodosus*	AF357202	113,193	19
Avermectin	*S. avermitilis*	AB032367	64,957	13
Erythromycin	*Sacc. erythraea*	AY661566	32,299	7
Megalomycin	*M. megalomicea*	AF263245	47,981	7
Niddamycin	*S. caelestis*	AF016585	41,097	8
Nystatin	*S. noursei*	AF263912	123,580	19
Pikromycin	*S. venezuelae*	AF079138	37,948	7
Pimaricin	*S. natalensis*	AJ278573	84,985	14
Rapamycin	*S. hygroscopicus*	X86780	107,379	15
Rifamycin	*A. mediterranei*	AF040570	109,528	11
Spinosad	*Sacc. spinosa*	AY007564	80,161	11
Tylactone	*S. fradiae*	U78289	43,280	8
NRPS				
Actinomycin	*S. chrysomallus*	AF134587	21,920	6
		AF047717		
		AF204401		
CDA	*S. coelicolor*	AL645882	40,653	11
Complestatin	*S. lavendulae*	AF386507	32,254	7
Pristinamycin	*S. pristinaespiralis*	NZ_ABJI00000000	33,340	7
Fengycin	*B. subtilis*	NC_000964	37,780	10
Lichenycin	*B. licheniformis*	NC_006322	26,178	7
Surfactin	*B. subtilis*	NC_000964	25,403	7
Tyrocidin	*B. brevis*	AP008955	33,711	10

coordinates to modules and domains with a prediction of the chemistry of any products.

Twelve well-characterized modular PKS clusters (Table 1) from *Streptomyces* and related genera were analyzed using the program. There were a total of 139 modules. The program predicted 1,954 pairs of recombination sites, which, as each recombination event produces two recombinants, would yield 3,908 recombinants. Most of the recombinants (81%) occurred between two KS domains (Fig. 1a; for details see Tables S1 and S2 in the supplementary material). The 12 clusters result in 66 cluster pairs, but the number of recombinants is very different for the different cluster pairs (see Table 2). In fact, 58% of the total recombinants are generated by just four cluster pairs, with more than 100 recombinants each. There were also five cluster pairs that generated no recombinants.

The cluster pair that generated the largest number of recombinants was amphotericin-nystatin, with 1,334 recombinants. The two clusters are closely related in sequence and gene organization [5, 8], so this is not surprising. The other cases with large numbers of recombinants also involved related clusters. In evolutionary terms, these are probably less interesting, because their capacity to generate new products will be similar to recombination events between two copies of the same cluster, which should occur at a much higher frequency. In contrast, recombination between less related clusters should produce novel cluster architectures. We, therefore, decided to concentrate on these cluster pairs.

If we remove the recombinants involving the related three clusters amphotericin, nystatin and pimaricin; and the related pair erythromycin and megalomycin, the remaining cluster pairs have an average of 26 recombinants per cluster pair. Of the recombinants, 94% were between two KS domains. The dominant role of KS domains has important implications for the functionality of recombinants.

Prediction of novel PKS products from recombinants

Although the analysis suggests that novel clusters should be produced by single cross overs between cluster pairs, it does not show whether they are likely to produce polyketide structures. Any recombination between domains of different types or between domains and linkers will disrupt modules so that the recombinant cluster is unlikely to be synthetically active. The dominant recombinant type is that

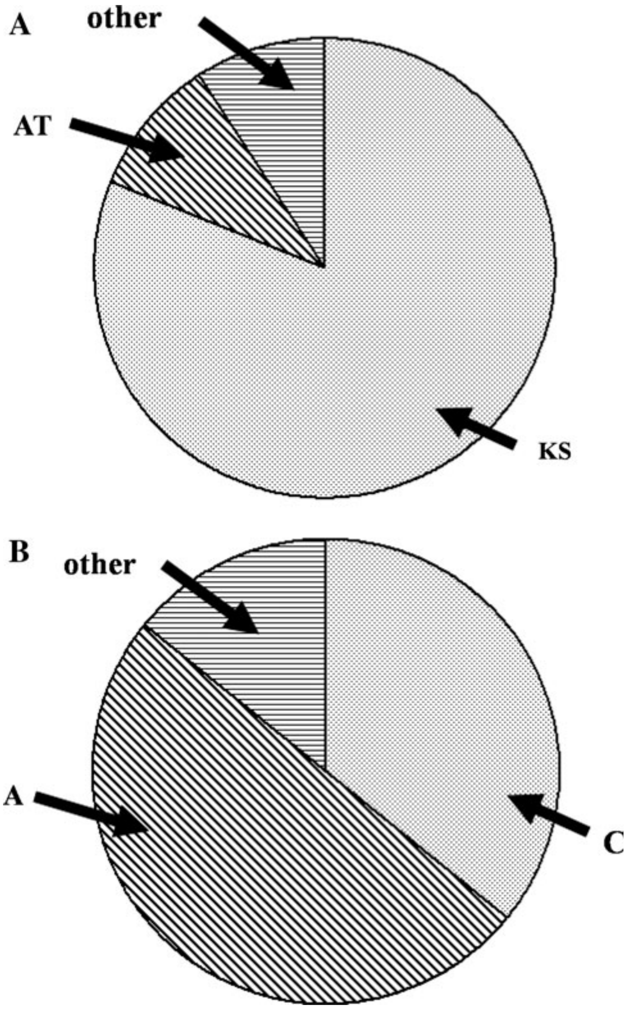

Fig. 1 Location of recombination sites in PKS and NRPS cluster pairs from *Streptomyces* and related genera. **a** Distribution of recombination sites in 1,954 recombination events in 66 PKS cluster pairs. **b** Distribution of recombination sites in 84 recombination events in six NRPS cluster pairs

arising from recombination between two KS domains. There is little evidence of KS domains contributing to specificity of the extension product; previously they were thought to determine the stereospecificity of condensation reactions, but more recent work has suggested that KR domains have this role [22, 33]. KS domains are highly conserved and there seems to be no reason for believing that a recombinant domain would be inactive. The KS domain is the first domain in each module. Thus, the specificity of the recombinant module should be identical to that of the module from the one parent, which contributes all the other specificity-determining domains.

Table 2 Total number (*upper right triangle*) of recombinants and number (*lower left triangle*) of recombinants predicted to produce a polyketide product

	amph	ave	ery	meg	nid	nys	pik	pim	rap	rif	spn	tyl
amph		30	34	8	80	1,334	52	420	20	44	4	26
ave	11		0	0	44	58	2	6	48	14	86	128
ery	8	0		78	8	36	22	16	0	16	2	14
meg	4	0	39		2	2	4	4	6	2	0	18
nid	40	40	4	2		98	38	52	4	14	0	20
nys	223	7	6	1	49		48	402	34	34	34	96
pik	1	1	0	2	19	14		32	2	32	2	8
pim	3	0	0	0	0	1	0		36	36	10	32
rap	1	14	0	0	2	1	0	10		12	40	26
rif	22	10	8	2	14	17	16	1	6		6	18
spn	0	7	0	0	0	10	0	1	6	3		74
tyl	13	28	7	18	20	48	4	0	6	18	37	

Apart from the activity of individual modules, the overall topology of the cluster is also important. It is necessary to have a loading domain (i.e., a starter module) and a termination domain, which may or may not have a thioesterase activity [23]. The individual polypeptides carrying the modules interact through specific docking domains [6] that ensure that the correct polypeptides are organized in the correct order in the PKS complex. There has been recent progress in predicting which pairs of docking domains interact with each other on the basis of their protein sequences [2]. In order to predict whether recombinant clusters are capable of producing a polyketide, we make the conservative assumption that docking domains interact if and only if they interact in a parental cluster. These conditions can be incorporated in an algorithm to predict whether a recombinant cluster will produce a polyketide:

1. There is a loading domain from one parental cluster which starts biosynthesis.
2. All modules needed for biosynthesis, from the loading domain up to the module in which recombination occurs, are present from this parent. As the recombinants are produced by a single cross-over, if these modules are present on several genes, the genes other than the one which contains the cross-over point will be complete and the protein products will interact correctly via their natural docking domains. The protein containing the cross-over junction will also be linked to the previous modules with a natural docking domain pair.
3. All subsequent modules required for biosynthesis, from the recombination point up until the last module (e.g., with a releasing thioesterase domain), are present and are derived from the second parent. The same considerations as for the modules derived from the first parent show that the modules from the second parent will be linked by their natural docking domain pairs if they occur in more than one gene.

If the generation of recombinants is viewed as a single step of a genetic algorithm, these criteria for the production of a polyketide can be viewed as a fitness criterion and it would be possible to examine the consequences of multiple rounds of recombination. The fitness criterion is a very stringent one as it is not excluded that recombinant clusters that fail one or more of the criteria will produce polyketides. For instance, alternative loading reactions might eliminate the need for a loading domain or nascent polyketides might be released from the synthase without a termination domain. Interaction between docking domain pairs from the two clusters might be able to bypass the criteria for which modules must be present. It should also be noted that the productive recombinants might contain extra modules (i.e., genes) from one or both parents that do not participate in the postulated biosynthetic reactions.

When the total number of recombinants from the 12 clusters were examined (Table 2), 21% satisfied the conditions for producing a product. However, when the four closely related cluster pairs with over 100 recombinants per cluster pair were removed, this rose to 35%. However, this still meant, on average, that there were nine recombinants per cluster pair that would yield a product. Detailed inspection of the results (Table 2) shows that the number of recombinants and the number predicted to produce polyketides varied widely between cluster pairs. The predicted recombinant polyketides usually have very different chemical structures than the parents. This is illustrated by a predicted polyketide structure (Fig. 2) from an in silico recombination between the niddamycin and erythromycin gene clusters. The recombinant polyketide cluster has seven modules, whereas the parents have eight and seven modules, respectively (Table 1). Comparison of the structures of the recombinant polyketide with the parents (Fig. 2) shows that such recombination events can generate significant chemical diversity.

Recombination between NRPS clusters

NRPS clusters are common in *Streptomyces* species (ca. 72% G + C) as well as in *Bacillus* species (ca. 40% G + C). We used four NRPS clusters from *Streptomyces* species and four NRPS clusters from *Bacillus* species (Table 1) with the recombination program. Detailed results are given in the supplementary material Table S3. As expected, no recombinants were detected between clusters from the different genera. The six *Streptomyces* NRPS cluster pairs gave a total of 168 recombinants, i.e., an average of 28 per cluster pair, which was comparable to the numbers obtained with the PKS clusters that were not closely related (average of 26 per cluster pair). When the sites of cross overs were examined (Fig. 1b), it was found that 36% of recombination events occurred between two C domains and 50% of recombination events involved two A domains. The same approach used for PKS clusters can be used to predict whether the recombinant NRPS clusters are productive. None of the 168 recombinant clusters were predicted to produce recombinants. Only one cluster pair from *Bacillus* species gave recombinants: surfactin and fengycin. There were two recombination events both involving a pair of A domains.

Detection of potential recombination genes and Chi sequences in *Streptomyces* genomes

No homologues of presynaptic proteins such as RecB or AddA have been detected in *Streptomyces* [30]. The protein sequences of four *uvrD*-family DNA helicases (UvrD and RecB from *Escherichia coli*, PcrA and AddA from *Bacillus subtilis*) were used for BLAST [1] searches against the *S. coelicolor* A3(2) proteome. Each query detected the same four potential helicase proteins coded by the genes with locus numbers SCO4797, SCO5183, SCO5184 and SCO5188. BLAST searches showed that all four genes are also highly conserved in other *Streptomyces* species (e.g., *S. avermitilis* and *S. griseus*), so they are candidates for a subunit of a RecBCD/AddAB-like enzyme that could interact with a Chi site. *Mycobacterium tuberculosis* has a *recBCD* operon [27]. Therefore, it might be expected that it would lack any helicase involved in recombination in *Streptomyces* species. However, when the deduced proteome of *M. tuberculosis* (genome sequence accession number AL123456) was used for BLAST searches, there were four *uvrD*-family helicases in addition to RecB. Each of them showed a closest match to one of the *S. coelicolor* loci: Rv0949 (60% amino acid identity with SCO4797), Rv3202c (40% with SCO5187), Rv3201c (39% with SCO5184), Rv3198c (51% with SCO5188). In both cases, there is one separated gene and three close to each other.

Chi sequences, which stimulate recombination by interacting with specific helicases, usually show an asymmetric distribution between the DNA strands with respect to the direction of replication forks [35]. In *Streptomyces* species, the replication origin *oriC* lies close to the middle of the chromosome and the two replication forks proceed to the chromosome ends, making assignment of genomic sequences to the leading and lagging strands easy. A program was written in Perl using the BioPerl package to search for oligonucleotides of 5–8 bases, which have an asymmetric distribution between leading and lagging strand. There are 2,266 copies of the 7-mer GGGGGAG in the genome of *S. coelicolor* A3(2) (one copy every 3.8 kb on average) with 80% of the copies on the leading strand. This distribution was maintained throughout the whole chromosome (Fig. 3), that results in a strand switch in

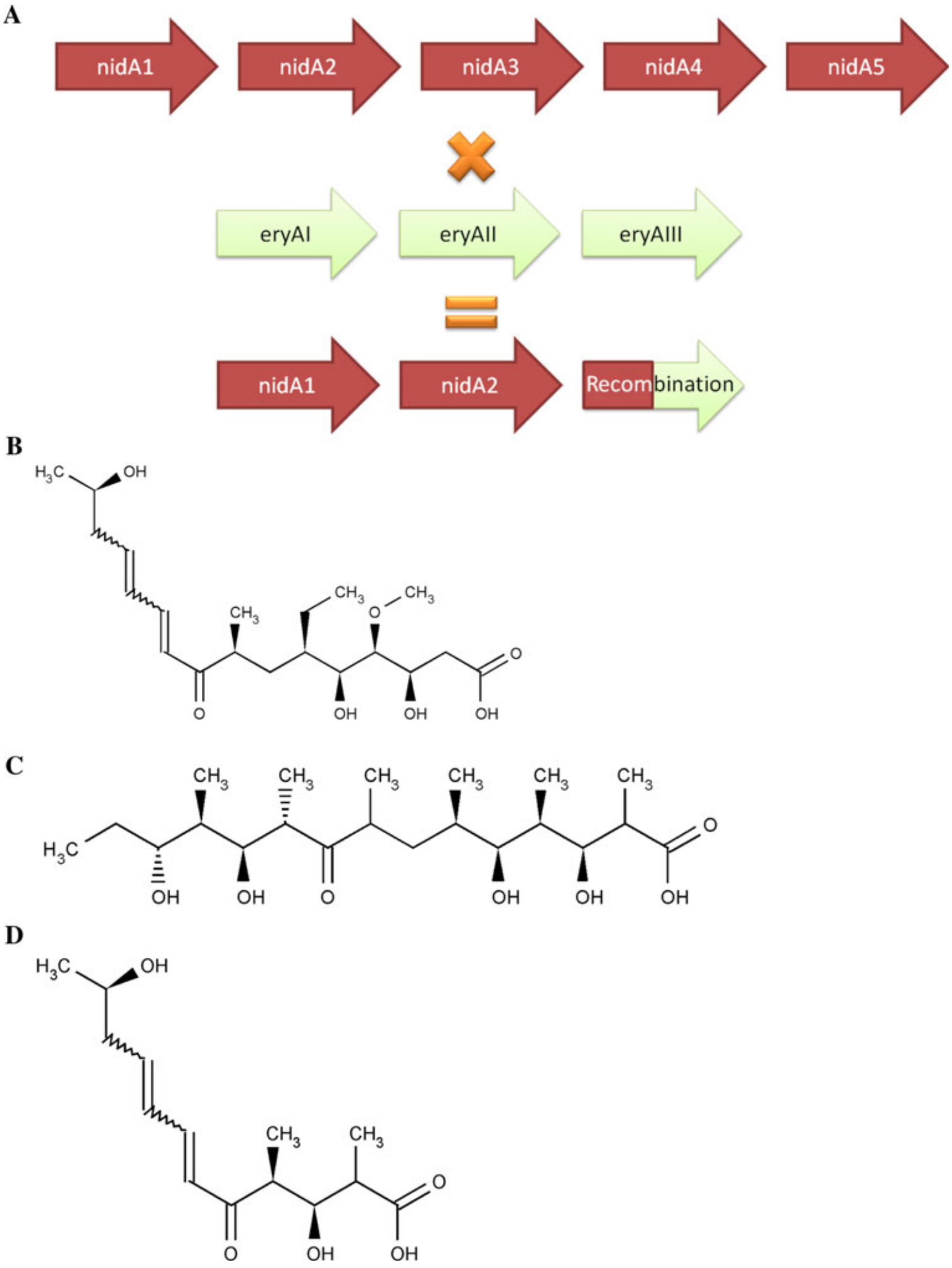

Fig. 2 Example of a predicted polyketide product arising from recombination between the niddamycin and erythromycin clusters. Recombination occurs between the KS domains of niddamycin module 5 (in gene *nid*A3) and erythromycin module 5 (in gene *ery*AIII). **a** Illustration showing that the recombinant is derived from the *left end* of the niddamyin cluster and the *right end* of the erythromycin cluster. The linear polyketide backbones of the parent molecules (**b** niddamycin, **c** erythromycin) and the recombinant product (**c**) are shown

preference for the sequence at *oriC*. Other 7-mers also showed a bias, but this was much less extreme: the next ten most extreme biases were 65–72%. The sequence GGGGGAG also showed a high strand bias in the chromosome sequences of *S. avermitilis* (data not shown).

The positions of the putative Chi sequence were compared with the chromosome annotation; 20% of the sequences were in non-coding sequences, which accounted for 11% of the annotated genome, so there was a small bias in favor of non-coding regions. The distribution of the putative Chi sites in 12 PKS gene clusters (Table 1) was also examined (data not shown). The short lengths of the clusters compared to chromosomes results in considerable statistical fluctuation for individual clusters, but the average bias for all cluster sequences (82%) and their frequency (one copy every 4.7 kb on average) are similar to the values for the chromosomes.

Discussion

Little is known about homologous recombination systems in *Streptomyces*. It seems likely that one of the four *uvrD*-family helicase genes might encode a RecBCD/AddAB-like

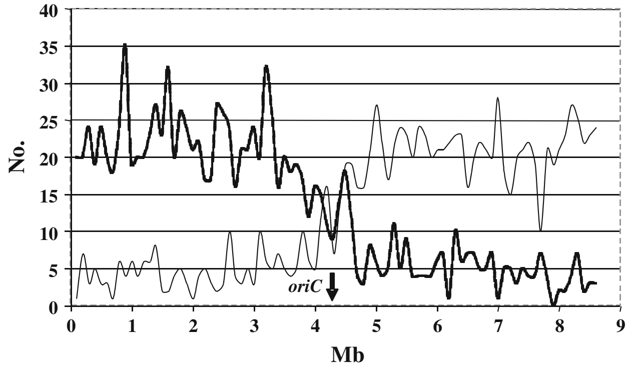

Fig. 3 Number of GGGGGAG sequences per 100 kb in the chromosomal DNA sequence of *S. coelicolor* A3(2) (*solid line*). The number of copies in the other strand is shown with the *shaded line* and the position of the replication origin (*oriC*) is indicated

recombination enzyme and this could be tested using a knock-out approach [20]. The main specificity for choice of recombination sites is probably determined during synapsis and would depend on the RecA protein. The presence of a Chi sequence would increase the frequency of successful recombinant formation post-synaptically. We identified a potential chi-sequence (GGGGGAG). Chi-sequences have been defined experimental for *E. coli* (GCTGGTGG), *B. subtilis* (AGCGG) and *Lactococcus lactis* (GCGCGTG), but show no obvious similarities. The activity of the putative chi-sequence could be tested experimentally as in *L. lactis* [4].

The recombination model includes three parameters determining the stringency of site selection. The model predicted recombination sites, but did not attempt to predict frequencies of recombination. Recombination frequencies decline as similarity of the sequences is reduced rather than showing an absolute presence or absence of recombination [11, 32]. Our recombination parameters were chosen to be non-stringent so that it is likely that we include many low-frequency recombination events. The recombination model was implemented as a module *CompGen* of the analysis program *ClustScan* [34, 38] so that it could link the DNA sequences to the synthetic order in clusters and derive the possible products of recombinant clusters.

Recombination was modeled between pairs of 12 well-characterized PKS-encoding gene clusters. The number of predicted recombinants varied very widely (Table 2). The highest number were generated by recombination between clusters that are closely related (amphotericin/nystatin/pimaricin and erythromycin/megalomycin). However, these recombinants are of less interest as they do not introduce new functions into clusters; recombination between two copies of the same cluster would be much more efficient and should produce a similar degree of chemical diversity. When these closely related cluster pairs were removed from the analysis, cluster pairs produced an average of 26 recombinants per cluster pair. Thus, it seems unlikely that constraints of homologous recombination will form a serious barrier to assembling clusters with different combinations of module types. However, in many cases there are only a small number of recombinants between clusters, e.g., the megalomycin-nystatin and the avermectin-pikromycin cluster pairs each only have two predicted recombinants (Table 2) despite using rather non-stringent recombination parameters. Constraints on recombination may result in segments of clusters being conserved. The approach used here should help in identifying parts of PKS clusters that are unlikely to recombine and lead to the identification of evolutionary conserved segments. Direct similarity-based methods to identify conserved segments in clusters are difficult to employ as there is extensive gene conversion between modules in PKS clusters (J. Zucko et al., unpublished data). This probably accounts for the observation made by several groups that most of the KS domains of a particular cluster tend to group together in phylogenetic trees [21].

Our analysis of PKS recombination used clusters from *Streptomyces* species as well as the genera *Amycolatopsis* (rifamycin), *Micromonospora* (megalomycin) and *Saccharopolyspora* (erythromycin, spinosad). It is striking (Table 2) that, in general, the frequency of recombinants between clusters from different genera is not lower than for clusters derived from *Streptomyces*. This is consistent with the idea that horizontal gene transfer is common for secondary metabolite clusters.

In some cases (e.g., the erythromycin cluster), the polypeptides of the PKS are all encoded by the same strand of the DNA molecule and are arranged in the biosynthetic order. If both clusters in a recombination event have such a genetic organization, recombinants generated by a single cross over will have the correct architecture to produce a polyketide. However, other clusters have a more complex organization (e.g., two of each of the polypeptides of the avermectin PKS are encoded on each of the DNA stands) and it is not obvious whether recombinant clusters will successfully produce polyketides. We developed an algorithm to test whether a contiguous set of biochemical steps were present making PKS production likely. This showed that for the less closely related clusters there were about 8 productive recombinants per cluster pair on average. Thus, a substantial proportion of recombinants should have a product that can be acted on by selection. This analysis does not consider the role of regulatory proteins that are needed for promoter activation. However, as recombination will produce two viable recombinant molecules in the cell, these proteins will be available. It is possible that even recombinant clusters that do not fulfill the stringent conditions for predicted productivity might produce polyketides. It is possible that an AT domain from an extender module could acquire a starter function; some loading domain AT domains are closely related to extender AT domains [34, 38]. If there were no natural termination module, it is still likely that polyketide chains will be released from the PKS with lower efficiency [23] and would probably undergo a cyclization.

Four NRPS clusters from *Streptomyces* species were analyzed with the recombination program and yielded a similar number of recombinants to PKS clusters. However, an examination of the recombination sites suggests that, despite a similar architecture, they might have different evolutionary paths to PKS-encoding gene clusters. In PKS encoding gene clusters, most recombinations occurred in KS domains, which would not lead to the generation of modules with novel activity. In contrast, many NRPS recombination events occurred in A domains, so that novel substrate specificities might be created. This might reflect the very large number of amino acid substrates (about 400 [7, 31]) naturally incorporated by different NRPS modules, compared to the limited number of natural CoA-ester PKS substrates (about 10 [9]). None of the recombinant clusters were predicted to be productive; this seems to contrast with the rate of productive clusters in PKS recombinants (about 30%), but an analysis of more NRPS clusters is necessary to see whether this is a general observation. Four NRPS clusters from *Bacillus* species were also analyzed and produced very few recombinants. More cases should be examined before concluding that this is a general property of *Bacillus* clusters. However, it would be consistent with the idea that the G + C-content plays an important role. When the G + C-content deviates from 50%, there is less choice of codons in protein coding regions [3] and this results in greater DNA sequence similarity between genes encoding homologous proteins. *Streptomyces* species (72% G + C-content) have a much larger deviation (22%) from 50% G + C-content than *Bacillus* species with a G + C-content of 40% and a 10% deviation. The mismatch repair system of *Bacillus* species would also be expected to radically reduce the recombination frequency [19], whereas actinomycete species do not contain homologues of the *mutS* and *mutL* genes that are responsible for this effect. Thus, there are at least three factors (linear chromosomes and plasmids, high G + C-content and lack of *mutSL* homologues), which would favor *Streptomyces* for the generation of new recombinant clusters. As well as this ability to generate new clusters, *Streptomyces* species are present in a soil environment, which is very competitive and applies a strong selection for the development of new secondary metabolites. This would help explain the large number and high diversity of modular biosynthetic clusters observed in *Streptomyces* species and related strains and why such organisms have been the most important source of secondary metabolites in screening programs. As homoeologous recombination can generate very diverse chemical structures (Fig. 2), it should be an important driving force in the chemical ecology of soil. Interactions occur between different groups of microorganisms (e.g., myxobacteria, fungi, slime moulds) and it is important to understand how each group generates new chemical diversity and the selection forces operating on the new chemical compounds.

The genetic tractability of *Streptomyces* strains makes construction of recombinants predicted to produce novel chemistries possible. This could either utilize recombination between clusters or the direct synthesis of predicted recombinant sequences. A very interesting question is whether such recombinant clusters would give a reasonable yield of the predicted novel product. Most genetic manipulations of modular PKS clusters (typically swapping of domains) result in extremely low yield of the recombinant polyketide. The reasons for this are unknown, but it is possible that these problems arise from mismatch in the junction sequences used to construct recombinants. If this were the case, utilization of homoeologous recombination sites, which are in regions of high sequence conservation, might alleviate the problem. Testing this hypothesis would be very interesting from the point of view of chemical ecology, where it is important to know how often recombinant clusters would synthesize novel compounds that would be subjected to selection. It would also show whether this approach had the potential to generate new chemical scaffolds of interest to the pharmaceutical industry.

Acknowledgments This work was funded by a cooperation grant of the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, Republic of Croatia (to J.C. and D.H.) and by the grant 058-0000000-3475 (to D.H.) from the Ministry of Science, Education and Sports, Republic of Croatia. Additional support was received from the Leverhulme Trust, Japanese Bio-Industry Association (to PFL).

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
2. Anand S, Prasad MV, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D (2010) SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res 38(suppl):W487–W496
3. Bibb MJ, Findlay PR, Johnson MW (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein coding sequences. Gene 30:157–166
4. Biswas I, Maguin E, Ehrlich SD, Gruss A (1995) A 7-base-pair sequence protects DNA from exonucleolytic degradation in *Lactococcus lactis*. Proc Natl Acad Sci USA 92:2244–2248

5. Brautaset T, Sekurova ON, Sletta H, Ellingsen TE, Strøm AR, Valla S, Zotchev SB (2000) Biosynthesis of the polyene antifungal antibiotic nystatin in *Streptomyces noursei* ATCC 11455: analysis of the gene cluster and deduction of the biosynthetic pathway. Chem Biol 7:395–403
6. Buchholz TJ, Geders TW, Bartley FE 3rd, Reynolds KA, Smith JL, Sherman DH (2009) Structural basis for binding specificity between subclasses of modular polyketide synthase docking domains. ACS Chem Biol 4:41–52
7. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G (2008) NORINE: a database of nonribosomal peptides. Nucleic Acids Res 36(Database issue):D326–D331
8. Caffrey P, Lynch S, Flood E, Finnan S, Oliynyk M (2001) Amphotericin biosynthesis in *Streptomyces nodosus*: deductions from analysis of polyketide synthase and late genes. Chem Biol 8:713–723
9. Chan YA, Podevels AM, Kevanya BM, Thomas MG (2009) Biosynthesis of polyketide synthase extender units. Nat Prod Rep 26:90–114
10. Chen CW, Huang CH, Lee HH, Tsai HH, Kirby R (2002) Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. Trends Genet 18:522–529
11. Datta A, Hendrix M, Lipsitch M, Jinks-Robertson S (1997) Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. Proc Natl Acad Sci USA 94:9757–9762
12. Denapaite D, Paravić Radičević A, Čajavec B, Hunter IS, Hranueli D, Cullum J (2005) Persistence of the chromosome end regions at low copy number in mutant strains of *Streptomyces rimosus* and *S. lividans*. Food Technol Biotechnol 43:9–17
13. Donadio S, Monciardini P, Sosio M (2007) Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. Nat Prod Rep 24:1073–1109
14. Egan S, Wiener P, Kallifidas D, Wellington EM (2001) Phylogeny of *Streptomyces* species and evidence for horizontal transfer of entire and partial antibiotic gene clusters. Antonie Van Leeuwenhoek 79:127–133
15. Fischbach MA, Walsh CT (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem Rev 106:3468–3496
16. Fischbach MA, Walsh CT, Clardy J (2008) The evolution of gene collectives: how natural selection drives chemical innovation. Proc Natl Acad Sci USA 105:4601–4608
17. Gravius B, Glocker D, Pigac J, Pandža K, Hranueli D, Cullum J (1994) The 387 kb linear plasmid pPZG101 of *Streptomyces rimosus* and its interactions with the chromosome. Microbiology 140:2271–2277
18. Hopwood DA (2006) Soil to genomics: the *Streptomyces* chromosome. Annu Rev Genet 40:1–23
19. Hranueli D, Cullum J, Basrak B, Goldstein P, Long PF (2005) Plasticity of the *Streptomyces* genome—evolution and engineering of new antibiotics. Curr Med Chem 12:1697–1704
20. Huang TW, Chen CW (2006) A *recA* null mutation may be generated in *Streptomyces coelicolor*. J Bacteriol 188:6771–6779
21. Jenke-Kodama H, Dittmann E (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. Nat Prod Rep 26:874–883
22. Keatinge-Clay AT (2007) A tylosin ketoreductase reveals how chirality is determined in polyketides. Chem Biol 14:898–908
23. Khosla C, Kapur S, Cane DE (2009) Revisiting the modularity of modular polyketide synthases. Curr Opin Chem Biol 13:135–143
24. Kinashi H, Shimaji-Murayama M, Hanafusa T (1992) Integration of SCP1, a giant linear plasmid, into the *Streptomyces coelicolor* chromosome. Gene 115:35–41
25. Kurosawa K, Ghiviriga I, Sambandan TG, Lessard PA, Barbara JE, Rha C, Sinskey AJ (2008) Rhodostreptomycins, antibiotics biosynthesized following horizontal gene transfer from *Streptomyces padanus* to *Rhodococcus fascians*. J Am Chem Soc 130:1126–1127
26. Mochizuki S, Hiratsu K, Suwa M, Ishii T, Sugino F, Yamada K, Kinashi H (2003) The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism. Mol Microbiol 48:1501–1510
27. Nakamura Y, Nishio Y, Ikeo K, Gojobori T (2003) The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. Gene 317:149–155
28. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
29. Pandza S, Biuković G, Paravić A, Dadbin A, Cullum J, Hranueli D (1998) Recombination between the linear plasmid pPZG101 and the linear chromosome of *Streptomyces rimosus* can lead to exchange of ends. Mol Microbiol 28:1165–1176
30. Rocha EP, Cornet E, Michel B (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. PLoS Genet 1:e15
31. Sattely ES, Fischbach MA, Walsh CT (2008) Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. Nat Prod Rep 25:757–793
32. Shen P, Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics 112:441–457
33. Starcevic A, Cullum J, Jaspars M, Hranueli D, Long PF (2007) Predicting the nature and timing of epimerisation on a modular polyketide synthase. Chembiochem 8:28–31
34. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D (2008) *ClustScan*: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res 36:6882–6892
35. Uno R, Nakayama Y, Arakawa K, Tomita M (2000) The orientation bias of Chi sequences is a general tendency of G-rich oligomers. Gene 259:207–215
36. Wei M, Wang S, Shang G (2010) Biosynthetic pathways and engineering for bioactive natural products. Curr Org Chem 14:1433–1446
37. Yamasaki M, Kinashi H (2004) Two chimeric chromosomes of *Streptomyces coelicolor* A3(2) generated by single crossover of the wild-type chromosome and linear plasmid scp1. J Bacteriol 186:6553–6559
38. Zucko J, Starcevic A, Diminic J, Elbekali M, Lisfi M, Long PF, Cullum J, Hranueli D (2010) From DNA sequences to chemical structures—methods for mining microbial genomic and metagenomic datasets for new natural products. Food Technol Biotechnol 48:234–242