BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 19 2007, pages 2543–2549
doi:10.1093/bioinformatics/btm381

Sequence analysis

Polyketide synthase genes and the natural products potential of *Dictyostelium discoideum*

J. Zucko[1,5], N. Skunca[1], T. Curk[2], B. Zupan[2,3], P.F. Long[4], J. Cullum[5], R.H. Kessin[6] and D. Hranueli[1,*]

[1]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia, [2]Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1001 Ljubljana, Slovenia, [3]Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA, [4]The School of Pharmacy, University of London, 29/39 Brunswick Square, London WC1N 1AX, UK, [5]Department of Genetics, University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany and [6]Department of Anatomy and Cell Biology, Columbia University, 630 West 168th Street, 10032 New York, USA

Received on May 15, 2007; revised on July 9, 2007; accepted on July 10, 2007

Advance Access publication July 27, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: The genome of the social amoeba *Dictyostelium discoideum* contains an unusually large number of polyketide synthase (PKS) genes. An analysis of the genes is a first step towards understanding the biological roles of their products and exploiting novel products.

Results: A total of 45 Type I iterative PKS genes were found, 5 of which are probably pseudogenes. Catalytic domains that are homologous with known PKS sequences as well as possible novel domains were identified. The genes often occurred in clusters of 2–5 genes, where members of the cluster had very similar sequences. The *D.discoideum* PKS genes formed a clade distinct from fungal and bacterial genes. All nine genes examined by RT–PCR were expressed, although at different developmental stages. The promoters of PKS genes were much more divergent than the structural genes, although we have identified motifs that are unique to some PKS gene promoters.

Contact: dhranueli@pbf.hr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The amoebae of *Dictyostelium discoideum* live in the soil and feed on a variety of bacteria and fungi. When the food is exhausted, the amoebae collect into mounds and then produce fruiting bodies. Many laboratories study the chemotaxis, the cell motility and the differentiation that are involved in fruiting body formation (Kessin, 2001). The genetics of the organism is well developed and it is possible to introduce genes and to knock-out genes. The genome sequence was recently completed showing about 12 500 genes in a relatively small AT-rich genome of 34 Mb (Eichinger *et al.*, 2005). The organism is exceptionally rich in polyketide synthases (PKS) with 43 putative genes spread singly and in clusters on all six chromosomes being reported in the initial annotation (Eichinger *et al.*, 2005). The only known polyketide product is differentiation inducing factor (DIF), which induces a particular subset of stalk cells during the complex development of the organism. The laboratory of Rob Kay characterized DIF and showed that its PKS is unusual in possessing a novel chalcone synthase domain (Austin *et al.*, 2006 and references therein). DIF has also been suggested to have mitochondrial uncoupling (Shaulsky and Loomis, 1995) and antiproliferation properties (Akaishi *et al.*, 2004; Kubohara *et al.*, 2003). There may well be further PKS genes involved in cell–cell communication, but it is also likely that some of them encode products to achieve competitive advantages in the soil (e.g. antibiotics). Polyketides are ubiquitous in nature and have been isolated from microorganisms, plants and invertebrates. They have found widespread use in the pharmaceutical, agrochemical and biotechnology industries. PKSs are large multienzyme protein complexes that contain a coordinated group of catalytic sites (Hranueli *et al.*, 2005). Type I PKSs are multifunctional proteins composed of all the active sites required for polyketide biosynthesis, which are contained in a series of domains in the proteins. Biosynthesis occurs as a stepwise process using simple carboxylic acid CoA esters as substrates. The minimal requirements for a PKS are the three domains: acyl carrier protein (ACP), acyltransferase (AT) and ketosynthase (KS). These are usually present in the order KS-AT-ACP in the protein and result in incorporation of a keto group. However, there are often one or more additional reduction domains present between AT and ACP: ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER), successively reduce the keto group to a hydroxyl group, a C=C double bond or a fully reduced product. Fatty acid synthases (FAS) belong to the PKS family and have all three reduction activities (Schweizer and Hofmann, 2004). When present, the reductive domains are in the order DH-ER-KR. There may also be a methyl transferase

*To whom correspondence should be addressed.

domain (MT) between DH and ER. It is common for inactive domains to be present, so that the presence of a domain does not prove that the corresponding activity occurs. The possibility to undergo none, one or more of these reduction reactions at each biosynthesis step contributes to the huge structural diversity seen in this class of natural products. The level of chemical complexity is further increased by incorporating stereo-isomers of different starter and extender units, as well as post PKS modifications, such as glycosylation, hydroxylation or methylation (Weissman and Leadlay, 2005). In bacterial systems, modular PKSs are common in which each biosynthetic step is carried out by a different module with its own KS-AT-ACP domains, which results in very large multi-modular proteins, e.g. for erythromycin biosynthesis (Khosla *et al*., 2007). In contrast, most fungal Type I systems are iterative so that a single module carries out several biosynthesis steps, e.g. for lovostatin biosynthesis (Schumann and Hertweck, 2006). In this article, we extend the analysis performed for this gene family by Eichinger *et al*. (2005). In collaboration with the curators at dictyBase (Chisholm *et al*., 2006), we have characterized the PKS genes of *D.discoideum* and shown that they are probably Type I iterative PKSs. However, many of the genes contain extra sequences that may be novel domains and there is an intriguing clustering of closely related genes in the chromosome. The PKS genes are differentially regulated, which is mirrored by divergence of the promoter regions.

2 METHODS

2.1 Identification and annotation of PKS genes

dictyBase version 2.5 (http://dictybase.org/; Chisholm *et al*., 2006) was used as a starting point for the analysis. The DNA sequences of the six chromosomes were translated using Transeq (Rice *et al*., 2000). BLAST searches used the NCBI server (http://www.ncbi.nlm.nih.gov/BLAST/; Altschul *et al*., 1990). Multiple alignments used the Clustal W service at EBI (http://www.ebi.ac.uk/clustalw; Thompson *et al*., 1994). Bacterial PKS domain sequences were obtained from the NRPS-PKS database (http://203.90.127.50/nrps-pks.html; Ansari *et al*., 2004). For profile analysis, HMMER version 2.3.2 (http://hmmer.janelia.org/; Eddy, 1998) and release 20 of the Pfam database (http://www.sanger.ac.uk/Software/Pfam/; Bateman *et al*., 2002) were used. Artemis (Rutherford *et al*., 2000) was used for annotation.

2.2 Phylogenetic analysis

KS and AT domain protein sequences were obtained for bacterial (Jenke-Kodama *et al*., 2005) and fungal PKSs (Ansari *et al*., 2004). Phylogenetic analysis was performed using the MEGA3 software package (Kumar *et al*., 2004). Uniform substitution rates at all sites were assumed and 100 or 500 replicates were used for bootstrapping. Trees were constructed with distance methods (neighbour-joining or minimal evolution) with two choices of distance model (Poisson correction or the Jones–Taylor–Thornton model). In addition, maximum parsimony was also used. Bacterial *fabF* and *fabD* genes or the chicken and human FAS genes were used as outgroups to root the trees. The sequences used are in the Supplementary Material.

2.3 Analysis of promoter regions

Motifs were detected with the program MEME (Bailey and Elkan, 1994). Motif combinations specific for PKS genes were detected using the program MAST (Bailey and Elkan, 1994). The selected motifs were matched with known motifs in the TRANSFAC (Wingender *et al*., 1996) database and sequence logos (Schneider and Stephens, 1990) were constructed. The program AlignACE (Hughes *et al*., 2000; Roth *et al*., 1998) was also used to find motifs.

2.4 PKS gene expression

Cell harvesting and RNA extraction were carried out with TRIZOL according to the recommendations of the manufacturer and as described in Van Driessche *et al*. (2005). Quantitative RT–PCR experiments were performed using syber-green real-time PCR on an Opticon system as described in Huang *et al*. (2006). Results were normalized to the transcripts of a constitutive gene, *IG7*. Gene-specific primer pairs were constructed for the following genes: *stlA* (*pks1*), *pks2*, *pks3*, *pks10*, *pks18*, *pks24*, *pks25*, *pks26*, *stlB* (*pks37*). The primer sequences are given in Supplementary Material.

3 RESULTS

3.1 Identification and domain structure of PKS genes

In dictyBase version 2.5, there were 46 sequences annotated as putative PKS genes as well as two genes annotated as putative fatty acid synthases. For our studies, the DNA sequence of each chromosome was translated in all six reading frames. Many of the putative PKS genes could be identified using BLAST with a standard KS domain (module 4 of erythromycin), but this did not show if a complete PKS gene was present. Therefore, HMM-profiles were constructed for the domains KS, AT, DH, ER, KR and ACP starting from well-characterized domains from bacterial modular PKS clusters and the sequences were analysed using the HMMER program package. The occurrence of typical PKS domains in the expected order adjacent to each other indicated the presence of a PKS gene. The newly identified domains from *D.discoideum* were used to refine the HMM-profiles to improve identification of PKS genes. This was particularly important for the DH domains, where the refined profiles recognized domains of the expected size in most of the genes, whereas the initial profile missed many DH domains. Methyl transferase (MT) domains are common in fungal PKSs (Schumann and Hertweck, 2006). MT domains were identified in many of the genes using an HMM-profile from the Pfam database (accession number PF08242). This initial analysis identified 45 putative PKS genes (*pks1–pks45*) (Supplementary Material). We did not rename the previously designated genes *stlA* (corresponding to *pks1*) or *stlB* (corresponding to *pks37*). This analysis showed considerable differences to the annotation in dictyBase version 2.5. In 14 cases, the previous annotation had recognized short sequences with resemblance to single PKS domains, but they did not have the structure of complete genes. In 12 cases, two different parts of a single PKS gene had been annotated as distinct PKS genes. There were 13 genes that had not been previously recognized as PKS genes and four new PKS genes. Although this analysis identified putative PKS genes, some sequences contained stop codons or had split domains into coding regions in different reading frames. This could either be because a pseudogene was present or because an intron had not been detected during the initial annotation. The predicted protein sequences were aligned with known PKS protein sequences using the CLUSTAL W program. In some

Fig. 1. Deduced protein structure of a typical PKS gene (*pks15*). The gene was not identified in the original genome annotation reported in dictyBase version 2.5. It is 9.453 kb in size and located on chromosome 2. In addition to the known domains, there are three amino acid regions of unknown function between the DH and MT domains, between the MT and ER domains and after ACP.

cases, it was clear that plausible intron splice sites would give rise to a protein sequence in good alignment with known domains. Forty of the 45 PKS genes had a structure compatible for expression. The five genes *pks4*, *pks11*, *pks12*, *pks20* and *pks43* are probably pseudogenes as they contain stop codons that cannot easily be explained by the presence of introns and in two cases also have an aberrant structure: *pks4* lacks an ACP domain and has incomplete ER and KR domains, while *pks12* has a 500 bp inversion with an adjacent 200 bp deletion. These changes have been coordinated with the curators of dictyBase, who have access to the original sequencing reads, and are now included in this online resource (Chisholm *et al*., 2006).

Among the 40 probably expressed genes, 2 were the previously described *stlA* (*pks1*) and *stlB* (*pks37*), which have chalcone-like domains. The other 38 genes all have the following set of predicted domains: KS, AT, ER, KR and ACP. Thirty seven of these genes have a DH domain of the correct length (Supplementary Material), whereas *pks21* has a short DH domain, which is probably not functional (94 instead of 154–187 amino acids in the other cases). Most genes (30/38) also contained an MT domain. Two genes that lacked the MT domain were *pks16* and *pks17*, which were suggested to be FAS. The DH and MT domains lie between AT and ER. However, in all 38 PKS genes there were also one or two substantial regions (286–838 amino acids) of no known function between AT and ER. This is shown for a typical gene (*pks15*) in Figure 1. Some of the sequences in this region are conserved between PKS genes. However, they do not give significant hits with BLAST to sequences in other organisms. Some of the sequences contribute to profiles in the Pfam-B database, but all members of the families are in the *D.discoideum* genome. Most known PKS proteins in other organisms end with the ACP domain. It was striking that 17 of the 38 predicted PKS proteins in *D.discoideum* had 288–501 additional amino acids after the ACP domain (Fig. 1 and Supplementary Material). These additional C-terminal regions did not resemble the chalcone-like domains at the C-terminals of *stlA* and *stlB*. Although there are conserved sequences between some of the genes, there was no detectable similarity to proteins in other organisms. The detailed structures of each PKS gene are given in the Supplementary Material. No Type II or modular Type I PKS genes were found. No non-ribosomally encoded peptide (NRPS) genes were found using BLAST or appropriate profiles.

3.2 Phylogeny of *D.discoideum* KS and AT domains and gene clusters

The amino acid sequences of the KS domains of the *D.discoideum* PKS genes were aligned with selected bacterial and fungal domains and used to construct a phylogenetic tree (Fig. 2) with neighbour-joining method using the Poisson correction distance model. Different tree construction methods were tested (Jones–Taylor–Thornton distance model, minimal evolution, maximum parsimony), but they did not result in significant changes in the tree (see Supplementary Material). The 45 *D.discoideum* PKS genes formed a clade (bootstrap value of 92%) distinct from the bacterial and fungal sequences. The genes *stlA* (*pks1*), *stlB* (*pks37*) and the two putative fatty acid synthase genes (*pks16* and *pks17*) were distant from the other 41 genes, which formed a clade (bootstrap value 91%). A phylogenetic tree was also constructed for the AT domains of the *D.discoideum* PKS genes (Supplementary Material). This tree showed an almost identical branching of the PKS genes compared to the KS tree. The *D.discoideum* AT-domains grouped with domains that incorporate C2 building blocks. Examination of the sequences showed that they contained C2-specific motifs (Haydock *et al*., 1995).

The protein sequences of the 38 *D.discoideum* PKS genes that are probably functional and possess all the domains KS, AT, DH, ER, KR and ACP were aligned with selected bacterial and fungal sequences that posses all the domains and used to construct a phylogenetic tree. The use of different tree construction methods (as for the KS trees) did not result in significant changes in the tree (see Supplementary Material). The *D.discoideum* PKS genes formed a clade distinct from the bacterial and fungal sequences. The phylogeny of the whole genes was little different from that of the KS and AT domains alone.

The genes are distributed over the six chromosomes. However, many of the genes are clustered. There are 10 pairs of genes (*pks11*/*pks12*, *pks16*/*pks17*, *pks22*/*pks23*, *pks24*/*pks25*, *pks27*/*pks28*, *pks33*/*pks34*, *pks35*/*pks36*, *pks38*/*pks39*, *pks40*/*pks41* and *pks42*/*pks43*), which are not only adjacent on the chromosome, but which are very closely related in sequence as shown by the phylogenetic trees (Fig. 2). There are also clusters of three (*pks19*/*pks20*/*pks21*), four (*pks29*/*pks30*/*pks31*/*pks32*) and five (*pks5*/*pks6*/*pks7*/*pks8*/*pks9*) PKS genes; in these cases, the KS and AT sequences of members of the clusters are always very closely related. Although some clusters contain genes of similar structure (e.g. *pks22*/*pks23*) there are also cases where the domain structure is different (e.g. *pks24* lacks an MT domain that is present in *pks25*). There are also 13 genes that do not appear to be clustered and these do not have any other PKS genes that are very closely related to them (Fig. 2). Four of the probable pseudogenes occur in clusters. The cluster *pks11*/*pks12* consists of two pseudogenes. As *pks43* seems to be a pseudogene, the cluster *pks42*/*pks43* would contain only one functional gene. *pks20* seems to be a pseudogene, but *pks19*/*pks21* could still form a functional cluster.

Most of the introns in the PKS genes occur in the KS domains. In order to compare the positions of introns in different genes, the deduced protein sequences were aligned to

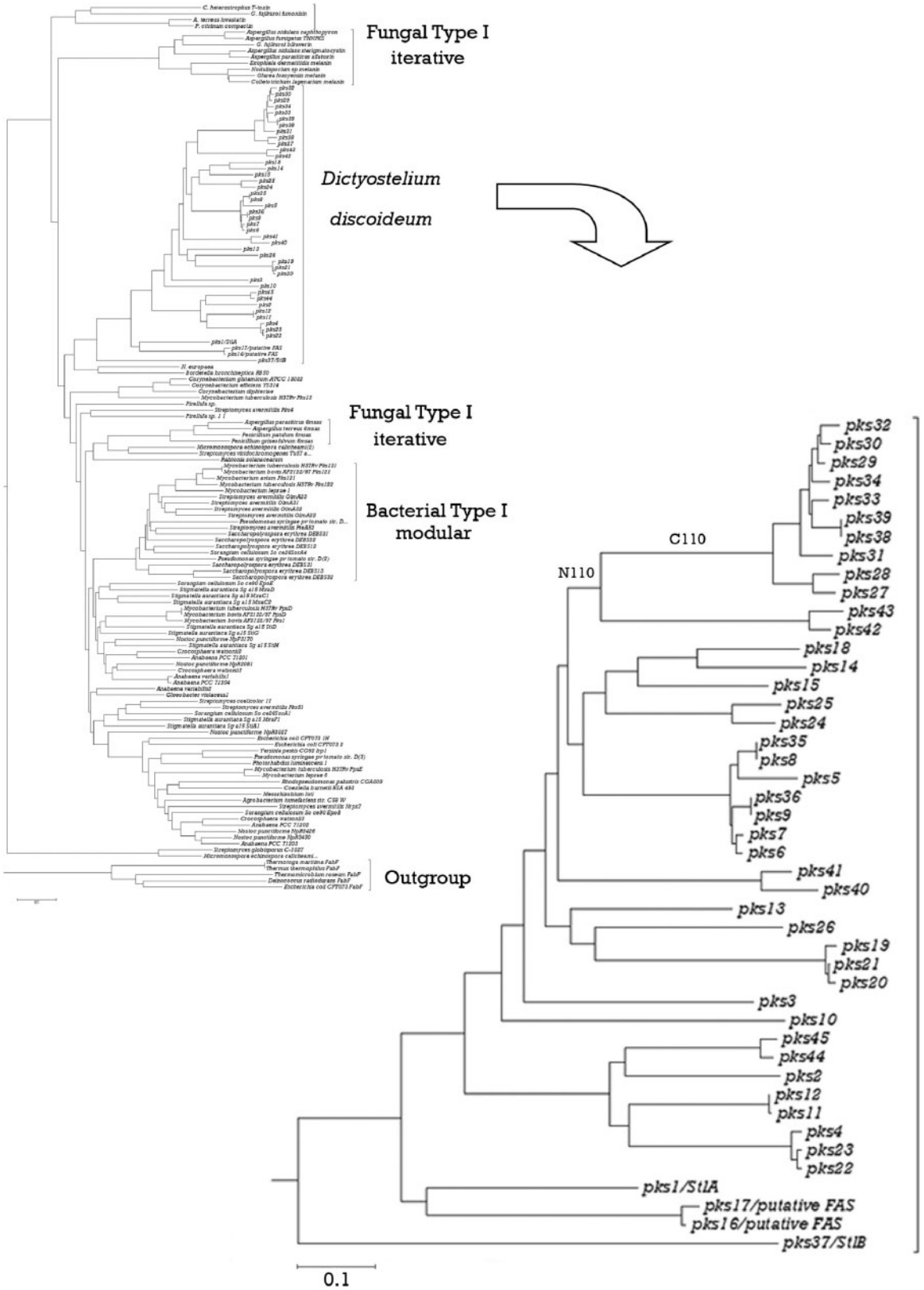

Fig. 2. Phylogenetic tree of the amino acid sequences of KS domains from bacterial, fungal and *D.discoideum* PKSs. The *D.discoideum* sequences form a distinct clade (bootstrap value 92%) which is expanded in this figure. Clades carrying the introns N110 and C100 are indicated. The scale is percentage of amino acid distance.

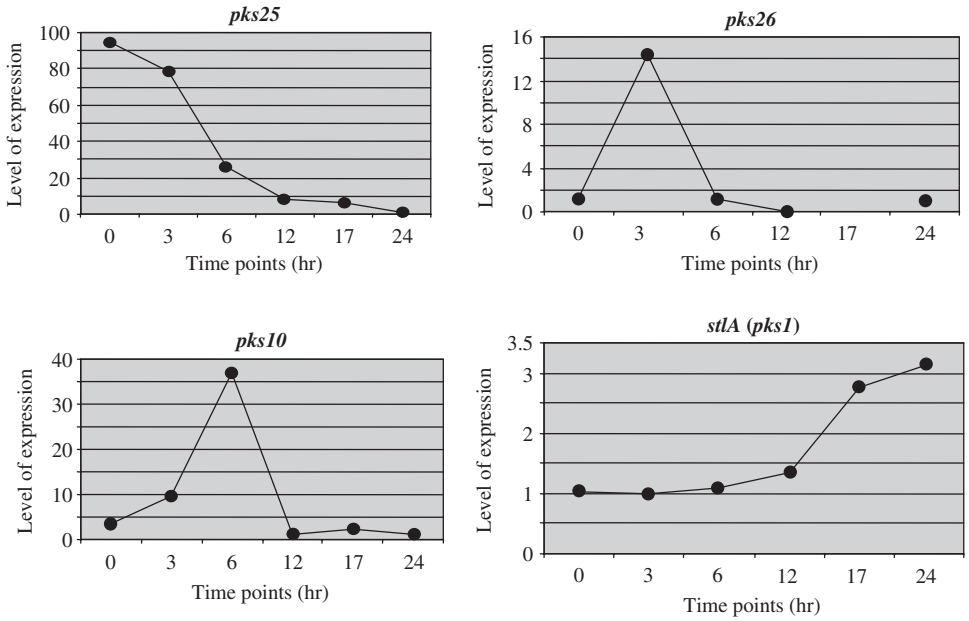

Fig. 3. Quantitative time course analysis of the gene expression of four PKS genes using real time PCR. The transcript levels were normalized using the constitutively expressed gene *IG7*.

Pfam HMM-profiles for the N-terminal (accession number PF00109.16) and C-terminal (accession number PF02801.12) parts of KS. The amino acid number of the profile at the intron position was used to localize the intron. Four intron positions were each found in 10 or more genes. The 10 PKS genes with an intron at position C100 form a clade (Fig. 2; bootstrap value 99%); the intron at position N205 shows an identical distribution except that an intron is also present at this position in *pks2*. The intron at amino acid positions N110 was present in all genes which contain C100 as well as *pks42* and *pks43*; these genes are also a clade (bootstrap value 47%). The intron C248 is present in 34 of the PKS genes. Eight of the genes which lack this intron are more distant in the KS tree (*stlA*, *stlB*, *pks16*, *pks17*, *pks11*, *pks12*, *pks44* and *pks45*). However, it is striking that it is also missing from three genes, whose neighbours all possess the intron (*pks10*, *pks15* and *pks19*), which suggests that the intron has been lost in these cases.

3.3 The transcription of PKS genes

Examination of expressed sequence tags in dictyBase version 2.5 indicated that most of the PKS genes (32/45) were transcribed. Two of the probable pseudogenes (*pks4* and *pks20*) have ESTs. Although the presence of EST clones proves transcription of the genes, it gives little information about timing and level of expression. A number of PKS gene sequences exist on standard microarray chips (G.Shaulsky, personal communication). These data, which can be found in the dictyBase version 2.5, indicate that the genes are transcribed. However, the high similarity of some *Dictyostelium* PKS genes would cause extensive cross hybridization. As a result, there is little correlation of expression patterns with more discriminating quantitative RT–PCR experiments.

The kinetics of expression of 9 PKS genes were followed using real time PCR. The results were normalized to the transcript of a constitutive gene, *IG*7 (Huang *et al*., 2006). Starvation was initiated at the start of the time course. After ~6 h cells begin to aggregate and spores and fruiting bodies are formed after ~24 h. Expression patterns could be divided into four categories. Gene *pks18* was expressed during growth but mRNA disappeared within 3 h of starvation. Genes representing a second transcriptional class (*pks24*, *pks25*, *pks26*) were expressed during early development after 3 h of starvation, but their expression had fallen by 6 h. Within this class there were detailed differences in kinetics (e.g. *pks25*, *pks26*; Fig. 3). Levels of gene expression for a third transcriptional class (*pks2*, *pks3*, *pks10*, *stlB* = *pks37*), peaked after ~6 h (Fig. 3) as the cells began to aggregate. A final category represented by a single example (*pks1* = *stlA*) was transcribed preferentially in late development, up to the formation of spores and fruiting bodies after 24 h (Fig. 3). The absolute levels of transcription varied with the measured peak levels differing ~80-fold between the highly expressed gene *pks3* and the weakly expressed gene *stlA*. These results suggest that the PKS genes examined were differentially expressed during different growth stages of *D.discoideum*.

The differential expression of the different PKS genes ought to be reflected in their promoter sequences. In order to examine the probable promoters, a sequence of 1 kb upstream of the start codon was examined for each of the 45 PKS genes. Initial analysis showed that, although the sequences of the coding regions of the PKS genes are highly conserved and can easily be aligned with each other, the similarity usually breaks off upstream of the start codon. Little is known of promoter structure in *D.discoideum*, which are very AT-rich (>95%), so the upstream sequences of the PKS genes were analysed to

identify motifs corresponding to possible binding sequences for transcription factors. The MEME program (Bailey and Elkan, 1994) was used to find 800 motifs of 6–17 bp long. After filtering out very similar motifs, further analysis was carried out using the remaining 676 motifs. The number of conserved motifs was calculated for each pair of PKS genes and most pairs showed little similarity. However, five pairs of genes had very similar upstream sequences (*pks8*/*pks35*, *pks4*/*pks23*, *pks9*/*pks36*, *pks38*/*pks39* and *pks30*/*pks32*). Only the last two pairs belong to clusters and the members of other clusters do not show many conserved motifs. However, a single conserved binding site would be enough to allow coordinated expression. The results with the PKS upstream regions were compared to a set of genes that would be expected to have conserved promoter regions (29 actin genes) and to a set of 47 randomly chosen *D.discoideum* genes. The 1 kb upstream regions of these genes were analysed to identify possible conserved motifs. On average, conserved motifs occurred 33 times in each actin gene upstream region compared to 3 times for the random genes. For the PKS genes the motifs occurred 12 times on average so that the upstream regions were much more similar to each other than those of the random genes. A similar picture was obtained with the numbers of common motifs between pairs of promoters: for the actin gene pairs common motifs occurred 10.6 times on average, compared to 0.6 times for the random genes and 2.6 times for the PKS genes.

The MAST program (Bailey and Elkan, 1994) was used to look for the PKS motifs in the upstream regions of all identified (14102) *D.discoideum* genes. This was used to identify motifs and combinations of 2 or 3 motifs that were specific for PKS genes (present in at least 5 PKS genes and present in less than 30 non-PKS genes). This identified 6 single motifs, 16 motif pairs and 53 motif triples containing 28 different motifs. None of the six single motifs showed any similarity to known regulatory sequences in the TRANSFAC database (Wingender *et al*., 1996). The motifs present in each PKS promoter region and the matches with the TRANSFAC database are shown in the Supplementary Material. The program AlignACE, which is based on a Gibbs sampling procedure (Hughes *et al*., 2000; Roth *et al*., 1998), was also used to find motifs. The PKS promoter regions were analysed using a range of parameter values to mimic the parameter space considered by the MEME program. The 28 PKS-specific motifs identified with MEME were compared with those of the same length found with AlignACE by calculating the average Pearson correlation between base frequencies for all positions in the two motifs. The locations of the AlignACE motif with the highest correlation score were compared with those of the corresponding MEME motif. Twenty of the motifs showed the same locations in the promoter regions (median distance between the MEME and AlignACE motifs not more than 2 bp). Thus, most of the MEME motifs were also found by AlignACE. A comparison of the MEME and AlignACE motifs is shown in Supplementary Material.

4 DISCUSSION

Forty-five PKS genes (including five probable pseudogenes) were identified distributed between the six *D.discoideum* chromosomes. In comparison to the genome annotation in dictyBase version 2.5, 17 PKS genes were added and 20 suggested PKS genes were removed. It is not surprising that an automatic annotation system is relatively inefficient in recognizing genes encoding complex multi-domain proteins when introns are present. The use of HMM-profiles and alignment with known genes helped to identify probable introns. All the genes had a structure typical of Type I iterative PKS genes. In addition to known domains, there were substantial protein coding regions between the AT and ER domains and sometimes after the ACP domain. These may be additional domains of unknown function. Determining the activities of new domains will require an investigation of the chemical structures of the polyketide products as no similarities to known proteins were found.

Phylogenetic analysis of the protein sequences showed that the *D.discoideum* PKS genes formed a discrete group separate from fungal and bacterial sequences (Fig. 2). The chalcone-like genes (*stlA* and *stlB*) and the probable FAS genes (*pks16* and *pks17*; Eichinger *et al*., 2005), were more distant from the others. A very unusual observation is that the genes occur in clusters of 2, 3, 4 or 5 very similar genes. This suggests that the clusters arose from duplications after evolutionary separation from the other PKS genes. In other organisms, PKS genes are usually present in single copies and the proteins are probably homodimers. Although there are some fungal systems (e.g. lovostatin) in which there are two PKSs involved, these PKSs have distinct activity and differ a lot in sequence (Schumann and Hertweck, 2006), unlike the case in *D.discoideum*. It is tempting to speculate that gene pairs allow the formation of heterodimers, which could perhaps extend the biosynthetic repertoire of Type I iterative genes by dividing successive synthesis steps between the two polypeptides. However, it is not clear why this would occur so often in *D.discoideum* and not be observed in other species. It will be interesting to see if this pattern is repeated in other related organisms. This unusual feature of the PKS genes will make it very interesting to characterize their gene products and to see if all the genes in a cluster are needed for successful biosynthesis.

Many of the genes had ESTs in the database. Thus, most of the genes are transcribed and it is likely that most are also translated. Two of the probable pseudogenes also have EST clones. It is conceivable that *trans* complementation between domains occurs as has been reported in other PKSs (Simunovic *et al*., 2006). Each of the nine PKS genes tested showed characteristic kinetics of expression. Most of them were induced by starvation, which supports the idea that the PKS genes may be involved in activities such as signalling between cells or protecting the differentiating organism from competitors and predators. In contrast to strong conservation of the protein sequences, the regions upstream of the PKS genes showed little conservation. Potential transcription factor-binding motifs were identified, but more experimental data on expression levels and kinetics are necessary to narrow down the significant motifs.

No genes encoding Type I modular PKSs (Weissman and Leadlay, 2005) or Type II PKSs (Petkovic *et al*., 2006) were identified. Similarly no non-ribosomal peptide synthetase genes were found. The genomes of several other Dictyostelid species

are about to be sequenced, including *D.purpureum*, *Polysphondylium violaceum* and *D.citrinum* (Baylor College of Medicine/Rice University/Joint Genome Institute) and *D.mucoroides* (G.Gloeckner personal communication.). These genomes may reveal yet greater varieties of PKS or non-ribosomal peptide synthetase genes. The potential genetic diversity of natural products in Dictyostelid populations, which are ubiquitous in forest and cultivated soils, is very high. We infer from the fact that these large genes have maintained their ORFs in the face of genetic drift that they are critical to the survival of the organisms whether, as in the case of DIF, they are used as developmental signalling molecules or to control potentially harmful bacteria, fungi or nematodes.

ACKNOWLEDGEMENTS

We would like to thank Gad Shaulsky and Jessica Svetz of the Baylor College of Medicine for help with the real time PCR experiments. The work of the curators of the dictyBase, Pascale Gaudet, Karen Pilcher and Petra Fey was critical to the annotation of the PKS genes. This work was supported by grant TP-05/0058-23 (to D.H.) from the Ministry of Science, Education and Sports, Republic of Croatia, by a cooperation grant of the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, Republic of Croatia (to D.H. and J.C.) and by a stipendium of the DAAD (to J.Z.), by a grant from the Slovenian Research Agency (to B.Z. and T.C) and by a grant from Ad-Futura, Science and Education Foundation of the Republic of Slovenia, Public Fund (to T.C.).

Conflict of Interest: none declared.

REFERENCES

Akaishi,E. *et al.* (2004) Differentiation-inducing factor-1-induced growth arrest of K562 leukemia cells involves the reduction of ERK1/2 activity. *Eur. J. Pharmacol.*, **485**, 21–29.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ansari,M.Z. *et al.* (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**, 405–413.

Austin,M.B. *et al.* (2006) Biosynthesis of *Dictyostelium discoideum* differentiation inducing factor by a hybrid type I fatty acid–type III polyketide synthase. *Nat. Chem. Biol.*, **2**, 494–502.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bateman,A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Chisholm,R.L. *et al.* (2006) dictyBase, the model organism database for *Dictyostelium discoideum*. *Nucleic Acids Res.*, **34**, D423–D427.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eichinger,L. *et al.* (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43–57.

Haydock,S.F. *et al.* (1995) Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases. *FEBS Lett.*, **374**, 246–248.

Hranueli,D. *et al.* (2005) Plasticity of the *Streptomyces* genome – evolution and engineering of new antibiotics. *Curr. Med. Chem.*, **12**, 1697–1704.

Hughes,J.D. *et al.* (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.

Huang,E *et al.* (2006) bZIP transcription factor interactions regulate DIF responses in *Dictyostelium*. *Development*, **133**, 449–458.

Jenke-Kodama,H. *et al.* (2005) Evolutionary implications of bacterial polyketide synthases. *Mol. Biol. Evol.*, **22**, 2027–2039.

Kessin,R.H. (2001) *Dictyostelium – Evolution, Cell Biology, and the Development of Multicellularity*. Cambridge University Press, Cambridge, pp. 1–294.

Khosla,C. *et al.* (2007) Structure and mechanism of the 6-deoxyerythronolide B synthase. *Annu. Rev. Biochem.*, **76**, 195–221.

Kubohara,Y. *et al.* (2003) DIF-1, an anti-tumor substance found in *Dictyostelium discoideum*, inhibits progesterone-induced oocyte maturation in *Xenopus laevis*. *Eur. J. Pharmacol.*, **460**, 93–98.

Kumar,S. *et al.* (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinformatics*, **5**, 150–163.

Petkovic,H. *et al.* (2006) Genetics of *Streptomyces rimosus*, the oxytetracycline producer. *Microbiol. Mol. Biol. Rev.*, **70**, 704–728.

Rice,P. *et al.* (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Roth,F.P. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Rutherford,K. *et al.* (2000) Artemis: sequence visualisation and annotation. *Bioinformatics*, **16**, 944–945.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Schumann,J. and Hertweck,C. (2006) Advances in cloning, functional analysis and heterologous expression of fungal polyketide synthase genes. *J. Biotechnol.*, **124**, 690–703.

Schweizer,E. and Hofmann,J. (2004) Microbial Type I fatty acid synthases (FAS): major players in a network of cellular FAS systems. *Microbiol. Mol. Biol. Rev.*, **68**, 501–517.

Shaulsky,G. and Loomis,W.F. (1995) Mitochondrial DNA replication but no nuclear DNA replication during development of *Dictyostelium*. *Proc. Natl Acad. Sci. USA*, **92**, 5660–5663.

Simunovic,V. *et al.* (2006) Myxovirescin A biosynthesis is directed by hybrid polyketide synthases/nonribosomal peptide synthetase, 3-hydroxy-3-methylglutaryl-CoA synthases, and trans-acting acyltransferases. *Chem. Biochem.*, **7**, 1206–1220.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Van Driessche,N. *et al.* (2005) Epistasis analysis with global transcriptional phenotypes. *Nat. Genet.*, **37**, 471–477.

Weissman,K.J. and Leadlay,P.F. (2005) Combinatorial biosynthesis of reduced polyketides. *Nat. Rev. Microbiol.*, **3**, 925–936.

Wingender,E. *et al.* (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.