Protist, Vol. 159, 21—30, January 2008
http://www.elsevier.de/protis
Published online date 10 October 2007

Protist

ORIGINAL PAPER

Novel Insights into Evolution of Protistan Polyketide Synthases through Phylogenomic Analysis

Uwe John[a,1,1], Bánk Beszteri[a], Evelyne Derelle[b], Yves Van de Peer[c], Betsy Read[d], Hervé Moreau[b], and Allan Cembella[a]

[a]Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany
[b]Observatoire Océanologique, Laboratoire Arago, Unité Mixte de Recherche 7628, Centre National de la Recherche Scientifique-Université Pierre et Marie Curie-Paris 6, BP 44, 66651 Banyuls sur mer cedex, France
[c]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B—9052 Ghent, Belgium
[d]Department of Biological Sciences, California State University San Marcos, San Marcos, CA 92096-0001, USA

Submitted November 27, 2006; Accepted July 31, 2007
Monitoring Editor: Hervé Philippe

Polyketide synthase (PKS) enzymes are large multi-domain complexes that structurally and functionally resemble the fatty acid synthases involved in lipid metabolism. Polyketide biosynthesis of secondary metabolites and hence functional PKS genes are widespread among bacteria, fungi and streptophytes, but the Type I was formerly known only from bacteria and fungi. Recently Type I PKS genes were also uncovered in the genomes of some alveolate protists. Here we show that the newly sequenced genomes of representatives of other protist groups, specifically the chlorophytes *Ostreococcus tauri, O. lucimarinus,* and *Chlamydomonas reinhardtii,* and the haptophyte *Emiliania huxleyi* also contain putative modular Type I PKS genes. Based on the patchy phylogenetic distribution of this gene type among eukaryotic microorganisms, the question arises whether they originate from recent lateral gene transfer from bacteria. Our phylogenetic analyses do not indicate such an evolutionary history. Whether Type I PKS genes originated several times independently during eukaryotic evolution or were rather lost in many extant lineages cannot yet be answered. In any case, we show that environmental genome sequencing projects are likely to be a valuable resource when mining for genes resembling protistan PKS I genes.

Key words: evolution; fatty acid synthase (FAS); ketoacyl synthase (KS); polyketide synthase (PKS); protists; phylogenomics.

Introduction

Polyketides are a structurally diverse class of natural products derived from the polymerization of acetyl and propionyl subunits in a process

[1]Corresponding author; fax +49 471 4831 1425
e-mail ujohn@awi-bremerhaven.de (U. John).
[1]These authors contributed equally to this work.

doi:10.1016/j.protis.2007.08.001

similar to fatty acid synthesis. Such compounds are of pharmaceutical and biomedical interest because many of them have potent biological effects, such as antibiotics, anti-tumor compounds, natural insecticides and immunosuppressive agents (Staunton and Weissman 2001). Numerous functions in nature have been proposed for these secondary metabolites, ranging from chemical defense to complex cell communication. The presence of polyketides in bacteria, fungi and streptophytes has been known for decades, but their occurrence in protists has only recently been confirmed. Analysis of metabolic pathways has established that certain potent polyether biotoxins found in free-living marine dinoflagellates and haptophytes are entirely or partially derived via polyketide biosynthesis (MacKinnon et al. 2006; Wright and Cembella 1998).

The polyketide synthase (PKS) enzymes responsible for the synthesis of polyketides are large multi-domain complexes that structurally and functionally resemble fatty acid synthase (FAS) enzymes involved in lipid metabolism. FAS and PKS catalyze the sequential condensation of acyl units onto a growing carbon chain and both enzymes possess a similar set of functional domains: ketoacyl synthase (KS), acyl transferase (AT), ketoacyl reductase (KR), dehydratase (DH), enoyl reductase (ER), acyl carrier protein (ACP) (or phosphopantetheine attachment site [PP]), and thioesterase (TE). Whereas FAS is dependent upon the presence of the complete set of aforementioned functional units, the minimal structure of PKS requires only the ACP, KS and AT domains for the condensation reaction. The other domains (when present) catalyze the stepwise reduction of the initial carbonyl units (Hopwood and Sherman 1990).

Polyketide synthases are generally classified into three major structural sub-groups. Type I PKSs are large, highly modular proteins, whereas Types II are aggregates of monofunctional proteins. Type I PKSs are among the largest known proteins encoded by a single open reading frame (ORF), encompassing as many as 12,000 amino acid residues. These enzymes include several modules, some of which are responsible for chain elongation while others catalyze the associated reduction steps. In most bacteria, each module directs one round of chain extension and post-condensation modification to generate non-aromatic polyketides. In fungi and some bacteria, each module/enzyme of Type I PKS is used iteratively, yielding either aromatic or non-aromatic compounds. By comparison, Type II PKSs are multi-protein complexes whereby the individual enzymes are used iteratively for each cycle of chain extension. These Type II complexes are found exclusively in bacteria for synthesis of aromatic polyketides. The Type III PKSs, also known as chalcone synthases, are homodimeric and function iteratively as condensing enzymes. Their distribution was believed to be essentially restricted to streptophytes, within which they employ unusual starter units to act directly on acyl-CoA thioesters, independently of ACP. Recent microbial genome sequencing, however, has revealed additional Type III PKSs in bacteria, most of which are of unknown function (Gross et al. 2006; Moore and Hopke 2001). In their study of bacterial Type I PKS evolution, Jenke-Kodama et al. (2005) concluded that FAS and PKS passed through a long joint evolutionary process with the modular PKS type arising from bacterial FAS and primary iterative PKS.

Our attention in this phylogenomic analysis focused on the origin and evolution of the biosynthetic genes for Type I PKS. These genes are well documented in bacteria and fungi, but have also recently been found in the genome of the apicomplexan parasite *Cryptosporidium parvum* (Zhu et al. 2002). Fragments of putative Type I PKS genes have also been identified in free-living marine protists belonging to the dinoflagellates and haptophytes, which are also known to produce polyketide-derived biotoxins (Cembella and John 2006; MacKinnon et al. 2006; Snyder et al. 2005). However, the evidence for Type I PKS genes from these marine protists has been questioned because of the difficulties in maintaining true axenic cultures (the absence of both free-living and endosymbiotic bacteria), thereby raising suspicions the PKS genes identified could be either of protistan or bacterial origin. Despite strong circumstantial evidence of their protistan origin, such as recent in situ hybridization experiments that have localized Type I PKS sequences to the nuclear genome of the toxic polyketide-producing dinoflagellate *Karenia brevis* (Snyder et al. 2005), large genome fragment sequencing is necessary to confirm their eukaryotic origin.

In contrast to the aforementioned distribution of putative and confirmed Type I PKS genes, no such genes have been confirmed in any other eukaryotic genome. The apparent disparate distribution of Type I PKS genes among the lineages of the eukaryotic evolutionary tree raises the question of the origin of these genes.

The increasing availability of eukaryotic genome sequences from various lineages initiated our interest in a broad survey and phylogenomic analysis of Type I PKS genes in diverse organisms. We screened most published as well as several ongoing eukaryotic genome sequence projects for candidate Type I PKS genes. Besides the included genome projects, data from several EST projects on protists are available or are in progress. We chose not to include these data sets in our analyses because the mostly short sequence fragments were not a reliable basis for sequence analysis. Furthermore, because of the potential problem of bacterial contamination, the eukaryotic origin of sequences from such projects is much more uncertain than in the case of complete or near complete genome sequences. We then established the phylogenetic relationships of the putative PKS I sequences found in our genome screen and compared their primary structure to known representatives of this gene family. This genomic analysis has provided novel insights into the distribution and evolutionary history of these key enzymes in secondary metabolism.

Results and Discussion

Distribution of Polyketide Synthases in Protists

Screening most available eukaryotic genomes from published and ongoing genome sequencing projects for candidate Type I PKS genes (Table 1 and Supplementary Table S1) showed that sequences with high similarity to Type I PKS genes were only present in a few lineages and

Table 1. List of eukaryotic genomes analyzed in this study for PKS type I genes.

Species	Taxonomy	Source	Abbreviations	PKS I type
	Chromalveolates			
Cryptosporidium hominis	Apicomplexa	COGENT		Yes*
Cryptosporidium parvum	Apicomplexa	COGENT	EAK87820_cParvum	Yes
Eimeria tenella	Apicomplexa	SANGER	dev_EIMER	Yes
Toxoplasma gondii	Apicomplexa	SANGER	Toxop_KS	Yes
Plasmodium falciparum	Apicomplexa	COGENT		No
Theileria parva	Apicomplexa	TIGR		No
Tetrahymena thermophila	Ciliophora	TIGR		No
Perkinsus marinus	Perkinsea	TIGR		No
Emiliania huxleyi	Haptophyta	JGI[#]	Ehux_contig	Yes
Thalassiosira pseudonana	Stramenopiles	JGI		No
Phaeodactylum tricornutum	Stramenopiles	JGI		No
Phytophthora sojae	Stramenopiles	JGI		No
Phytophthora ramorum	Stramenopiles	JGI		No
	Excavata			
Leishmania major	Kinetoplastida	SANGER		No
Trypanosoma brucei	Kinetoplastida	SANGER		No
Trypanosoma cruzi	Kinetoplastida	TIGR		No
Naegleria gruberi	Heterolobosea	JGI		No
	Unikonts			
Monosiga brevicollis	Ophistokonts	JGI		No
	Plantae			
Cyanidioschyzon merolae	Rhodophyta	COGENT		No
Galdieria sulphuraria	Rhodophyta	MSU		No
Chlamydomonas reinhardtii	Chlorophyta	JGI	chlamyV3	Yes
Ostreococcus lucimarinus	Chlorophyta	JGI	opPKS	Yes
Ostreococcus tauri	Chlorophyta	Genopole Languedoc-Roussillon	KSotauri	Yes

were dispersed throughout the eukaryotic tree (our criteria for "high similarity with PKS I genes" were: regions showing sequence similarity ($e < 10^{-5}$) to different PKS I domains in close proximity within a contiguous ORF or at least in a contiguous chromosomal region).

Among lineages of Plantae, we found candidate sequences in the genomes of unicellular green algae (Chlorophyta), namely, *Chlamydomonas reinhardtii*, and in two recently sequenced *Ostreococcus* species, *O. tauri* (Derelle et al. 2006) and *O. lucimarinus* (Palenik et al. 2007). Both *Ostreococcus* species have a compact genome of approximately 13 Mbp (Derelle et al. 2002) and both genomes contain three large PKS genes. The corresponding polyketide products and their potential function are unknown, but the fact that the three PKS ORFs account for up to 1.5% of the genome indicates that they likely have an important function in this small genome (Keeling and Slamovits 2005). In any case, the presence of Type I PKS genes among green algae was surprising because such sequences were not found in the genomes of other Plantae lineages, neither in streptophytes nor in the red algae (Rhodophyta), *Cyanydioschyzon merolae* and *Galdieria sulphuraria* (Table 1).

The alveolates, which comprise the three sublineages Apicomplexa, dinoflagellates and ciliates, also have representatives possessing Type I PKS genes. Among Apicomplexa, Type I PKS and phylogenetically closely related FAS genes have been found in the parasites *Cryptosporidium parvum* (Zhu et al. 2002, 2004), *Eimeria tenella* (http://www.sanger.ac.uk/Projects/E_tenella/), and *Toxoplasma gondii* (http://www.sanger.ac.uk/Projects/T_gondii/). Similar sequence fragments can also been found in the genome of *Cryptosporidium hominis*. Nevertheless, the existence of Type I PKS genes is not universal among alveolates because these sequences were lacking in other species, including the parasites *Perkinsus marinus*, *Plasmodium falciparum* and *Theileria parva* and the free-living ciliates, *Tetrahymena thermophila* and *Paramecium* spp. Dinoflagellates, another important group of alveolates, appear to contain Type I PKS genes as indicated by several cDNA sequencing studies (Cembella and John 2006; Jaeckisch et al. 2007; Lidie et al. 2005; Snyder et al. 2003, 2005). In these cases, only short fragments were recovered and no full genome sequences are available for this group, therefore we did not include them in our analyses.

Among the stramenopiles, genomes of the oomycetes *Phythophthora ramorum* and *P. sojae*, and the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* apparently lack Type I PKS genes. The genome of another chromalveolate, the haptophyte *Emiliania huxleyi* was recently released (Version 1, 2006, bread@csusm.edu), but it has not yet been annotated. Long contiguous ORFs containing typical Type I PKS domains (KS, AT, ACP, KR, ER) are present on several scaffolds clearly corresponding to the *E. huxleyi* genome, based on their nucleotide composition, and not to bacterial contaminants, which had been removed from the assembly. This provided strong evidence that this species has at least one and perhaps several Type I PKS genes.

Of the three available genomes of Euglenozoa (*Trypanosoma brucei*, *T. cruzi* and *Leishmania major*), which are members of the Excavata, none exhibited PKS Type I sequences.

This patchy distributional pattern of PKS genes among members of the eukaryotic crown lineages is similar to findings for prokaryotes. In a study of 138 bacterial genomes, PKS genes could be identified in only 27 (21% of total), and none of the archaean genomes possess putative PKS genes (Jenke-Kodama et al. 2005).

Our further analyses focused on the beta-ketoacyl synthase (abbreviated as KS) domain, the most conserved domain of Type I PKS genes (Kroken et al. 2003). This domain has the greatest potential for revealing divergent homologes and thus provides the most informative basis for comparative and phylogenetic analyses. In phylogenetic analysis of the conserved KS domains of these newly discovered Type I PKS genes, we compared these gene sequences to known representatives from bacteria and fungi, and to the highly similar KS domains from metazoan FAS. These phylogenetic analyses (Fig. 1) recovered the clades known from previous phylogenetic analysis of this PKS domain (Kroken et al. 2003). The bacterial clades, the metazoan FAS, and Ascomycota KS clades (both reducing and non-reducing) were supported by bootstrap (BP) and posterior probability (PP) values (Fig. 1). The putative protistan PKS I KS sequences did not resemble any of these formerly described groups, but instead, clustered distinctly from them. The KS sequences from Apicomplexa (*C. parvum*, *T. gondii*, *E. tenella*) formed a well-supported monophyletic clade. The *Ostreococcus* Type I PKS sequences were spread into two clades-Chlorophyta Clade 1 and Clade 2. Chlorophyta Clade 1, consisting of 32 *O. tauri* (KSotauri), 21 *O. lucimarinus* (opPKS), and three *C. reinhardtii* KS sequences, grouped together with the Haptophyta

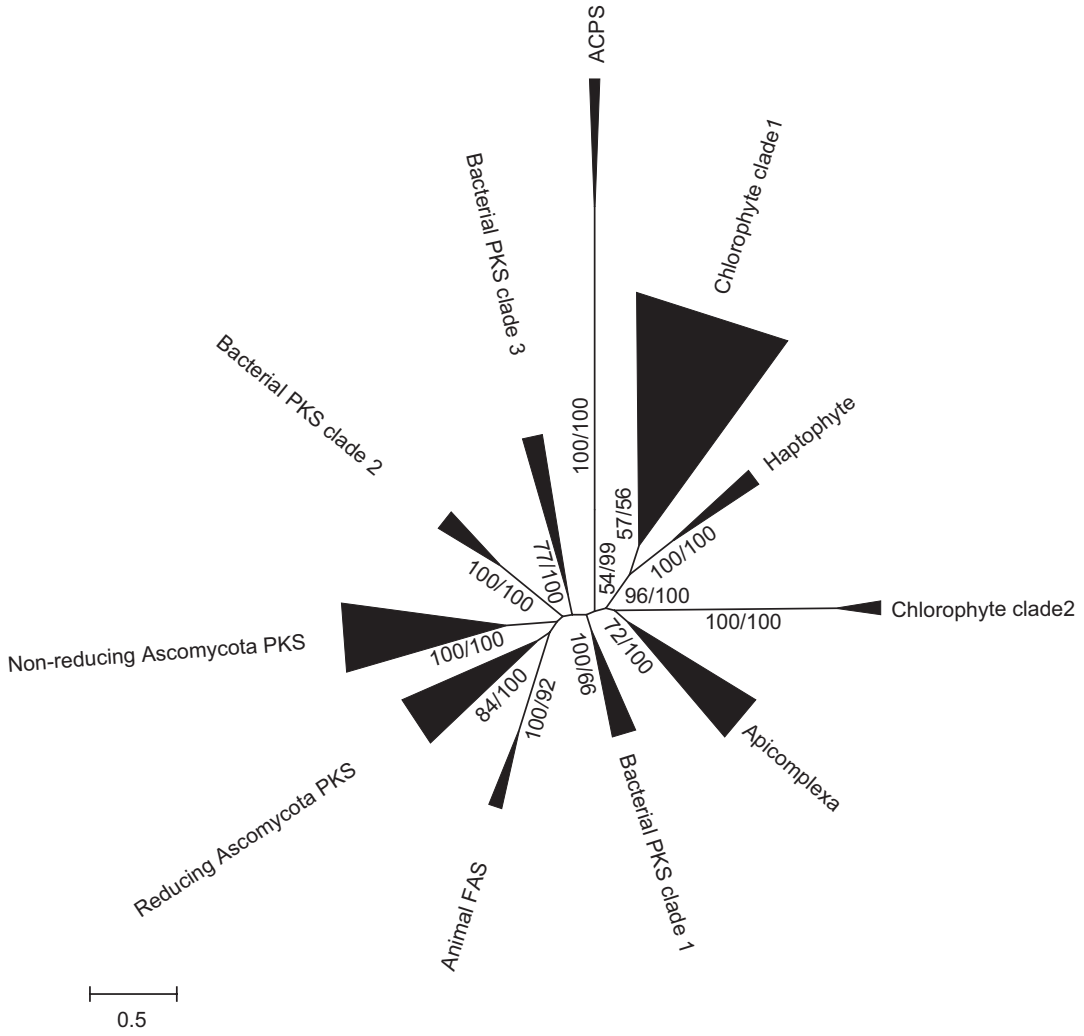

Figure 1. Maximum likelihood phylogenetic tree of the KS domain of Type I PKS sequences. Scale bar represents corrected evolutionary divergence. Numbers at the branches indicate bootstrap and posterior probability values (BP/PP).

clade (*E. huxleyi* KS sequences), whereas Clade 2, containing three *O. tauri* and one *O. lucimarinus* KS sequences was distinct. This grouping was surprising because chlorophytes are supposed to be a monophyletic group (Bhattacharya and Medlin 1998). Furthermore, within Chlorophyta Clade 1, sequences from *C. reinhardtii, O. tauri* and *O. lucimarinus* formed a relatively tight but mixed group, not reflecting the presumed long independent evolutionary history of these three taxa (Fig. 2). Whereas the monophyly of all described clades was well supported by BP and PP values, the relationships among them remain largely unresolved.

Considering the large divergence in the KS data set, we also tried to identify alignment positions at which substitution saturation has occurred using the program Asatura (van de Peer et al. 2002). We generated trees with different degrees of highly variable sites, but all tree topologies and bootstrap values remained basically unchanged by these manipulations (data not shown).

One of the most remarkable integrative results of our analyses was that the newly identified modular Type I PKS genes from protists differ from all hitherto known genes of this type. Whereas in bacteria the Type I PKS pathway frequently co-occurs with non-ribosomal peptide synthases (NRPS), regulating a second type of secondary metabolite pathway (Shen et al. 2001), such genes have not yet been found in protists. Furthermore, our analyses based on many currently available

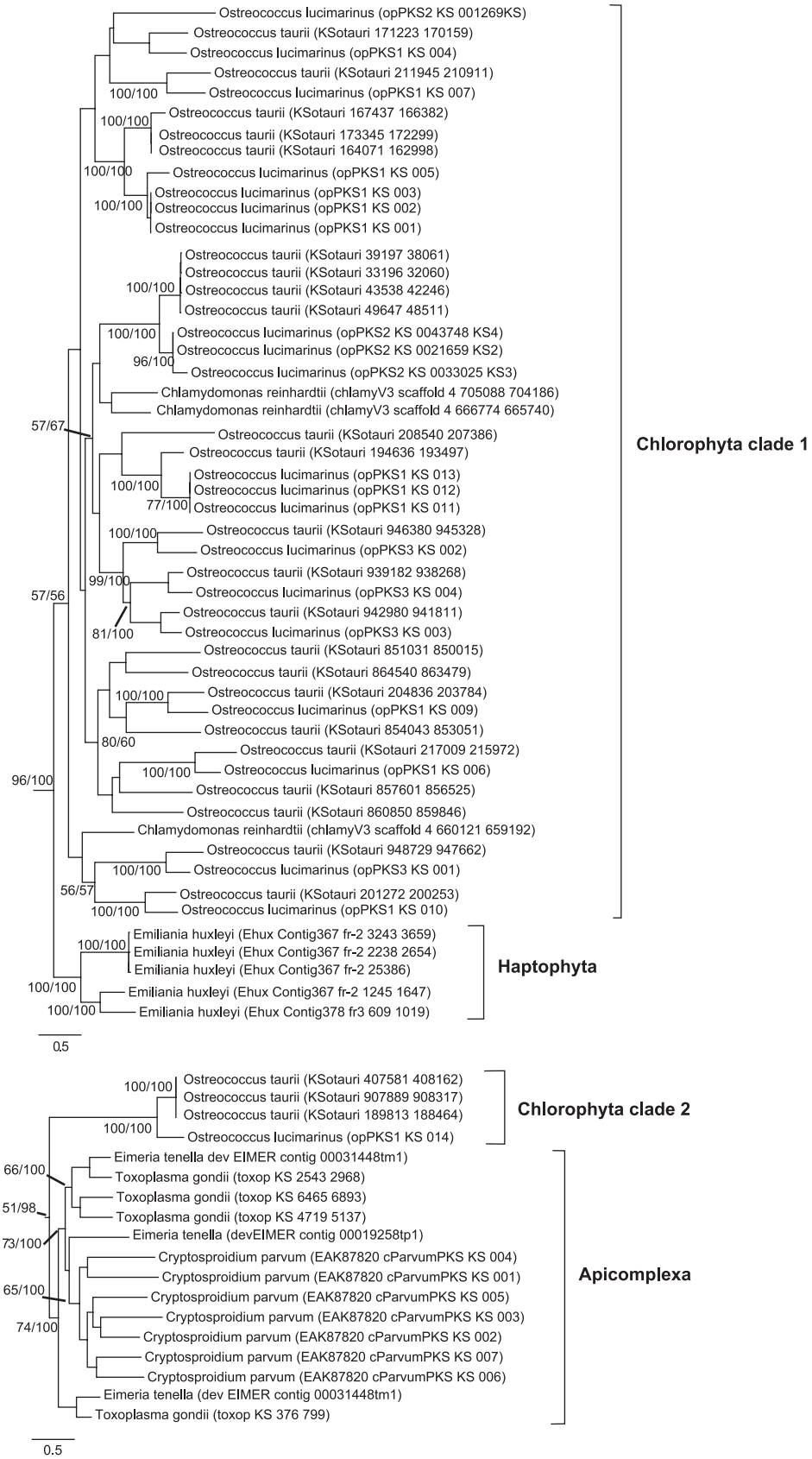

Figure 2. Maximum likelihood phylogenetic trees of the protistan clades. Corresponding taxon names can be taken from Table 1. Numbers at the branches indicate bootstrap and posterior probability values (BP/PP).

protistan genome sequences indicate that the occurrence of Type I PKS genes is much wider than traditionally thought. Our survey reveals that from the four large groups of eukaryotes for which genome sequences are currently available (unikonts, Plantae, excavates, and chromalveolates), three contain PKS I gene members (up to now, no Type I PKS was found in members of the Excavata). Nevertheless, these eukaryotic PKS genes do not group with any of the previously described clades (Kroken et al. 2003; Shen et al. 2001), but rather cluster separately in well-supported monophyletic clades that do not always fit well with the proposed phylogenetic relationships of their hosts (Baldauf 2003; Keeling et al. 2006).

Discrimination of Protistan from Bacterial and Fungal Beta-Ketoacyl Synthase Domains

The possibilities of a sequence-based discrimination of protistan PKS fragments was approached by using hidden Markov models (HMMs) of KS domain sequences. First, HMMs for bacterial, fungal and protistan KS domains were calculated based upon those sequences included in our alignments, i.e., 68 protist, 35 fungal and 19 bacterial sequences (for graphical representation of these profile models, see Supplementary Figures S1 and S2). Each sequence included in our alignment was then scored with the three models. As expected the best scores were obtained when sequences were compared to the model of their own group (i.e., protistan sequences with the protistan HMM, etc., Supplementary Table S2). The difference between the top scoring and second best model was always large (>120). The *e*-values from the best model were always at least seven orders of magnitude smaller than the second best model (when ignoring two *C. reinhardtii* sequences, this difference was always larger than 30 orders of magnitude). This suggests that it is possible to discriminate protistan from bacterial and fungal PKS I KS sequences by scoring protein sequences with these HMMs. Protistan KS sequences will conform better to the protistan model (higher scores/lower *e*-values) than to the other models.

To illustrate the potential of this approach in finding new candidate sequences by mining sequence databases, we screened the env_nr (environmental shotgun sequences) database of NCBI for sequences showing high similarity to KS domains of PKS I by tblastn searches, using as query a KS domain from *C. parvum* translated the high scoring segments of the first 100 hits, and scored them with the three HMMs (protistan, fungal and bacterial KS). Although the differences among the scores with different models were not always as well defined as in our original data set, we identified several candidate sequences which likely encode protistan KS domains from PKS I genes (see Supplementary Table S3). This suggests that environmental sequencing projects have a high probability of identifying other novel protistan PKS genes, despite the fact that most of these projects are aimed at the prokaryotic size fractions of environmental samples.

Evolution of PKS in Protists

The global eukaryotic phylogeny is still unresolved and thus the phylogenomic relationships among the groups must be cautiously interpreted. Furthermore, the evidence for monophyly of each eukaryotic super-group remains inconclusive and is still the subject of debate. With these caveats in mind, our results nevertheless highlight the irregular phylogenetic distribution of Type I PKS genes in eukaryotes. Whereas representatives containing Type I PKS are known for three above-mentioned groups (unikonts, Plantae, and chromalveolates), they also contain members that do not possess this gene.

The view that Type III PKS is characteristic for Plantae is now challenged by the finding of Type I PKS in all three available chlorophyte genomes (*Chlamydomonas reinhardtii*, *Ostreococcus taurii* and *O. lucimarinus*). Taking into account the grouping of the PKS sequences of *Chlamydomonas* and *Ostreococcus*, and the fact that those taxa diverged long ago, this would put the gene divergence event near the base of chlorophyte evolution. However, no Type I PKS genes were found in another lineage of the Plantae, i.e. in the Rhodophyta *Cyanydioschyzon merolae* and *Galdieria sulphuraria*, the only two red algae thus far sequenced or in progress. A simple explanation is that these genes have been lost during evolution of the rhodophytes. Nevertheless, both rhodophyte species that have been sequenced are unicellular and live in acidic hot springs, and are thus atypical. It is therefore important to confirm whether or not PKS genes are generally absent from rhodophytes, including species living in less extreme (e.g., temperate or tropical marine) habitats.

There are several possibilities to explain the irregular phylogenetic distribution of Type I PKS genes among protists. One possibility is that a

common ancestor of the lineages Plantae, unikonts and chromalveolates did indeed possess a Type I PKS gene, which was later lost among several of the descendant lineages (at the least five losses, based upon the rather limited taxon sampling available and depending upon the accuracy of the proposed phylogenetic relationships of the taxa included). One can speculate that these sequences were replaced by other types of PKS genes, such as PKS III in Plantae, and that they might have become unnecessary and then lost in several other groups.

An alternative scenario is that PKS I genes were obtained independently in several eukaryotic lineages, either through duplication and divergence from existing FAS genes in their genomes or via lateral gene transfer events. The phylogenetic relationships of the FAS and PKS genes found in *C. parvum* attest to an evolutionary origin of PKS I from FAS (Zhu et al. 2002). Examples for lateral gene transfer from bacteria into fungi were found by Kroken et al. (2003). In fact, the origin of Type I PKS from indigenous FAS in one lineage would not exclude a lateral gene transfer origin of PKS I in another group. Phylogenetic analyses have the potential to point towards the validity of these alternative explanations. Recent gene transfer events would be revealed by grouping transferred PKS I genes with those related to the organism of origin.

Our phylogenetic analyses showed the newly identified PKS I genes from protists did not group with any previously known PKS I group. Thus, our results do not provide an indication that the protistan PKS I genes have originated from relatively recent gene transfer events from bacteria or fungi, rather than early in the evolution of the lineages. Nevertheless, even if we preliminarily exclude lateral gene transfer as unlikely, our results do not allow us to state unequivocally which of the remaining two hypotheses is more likely valid—multiple losses or multiple, more ancient acquisitions of Type I PKS genes in diverse protistan lineages. The unexpected grouping of *E. huxleyi* with chlorophyte (instead of alveolate) KS might point to a common origin of *E. huxleyi* and green algal PKS, as distinct from that of alveolates. Yet we cannot exclude the possibility that the grouping of *E. huxleyi* with chlorophytes (and exclusion of alveolates) does not fully reflect the evolutionary relationships of these genes. Sequences of some hundreds of amino acids in length might not contain sufficient phylogenetic information to correctly resolve relationships on time scales comparable to the age of eukaryotes.

Perhaps the key to understanding the origins and evolution of the PKS genes in protists lies in the analysis of the functional significance of these genes and their role in secondary metabolism. The maintenance of PKS genes, which have no known obligate role in primary metabolism, over long evolutionary time suggests a persistent strong positive selection, but which is not necessarily well correlated with the general evolution of the taxa. For example, the selective pressure on secondary metabolite production among free-living marine flagellates of diverse phylogenetic lineages may be more similar than that exerted upon free-living forms of phylogenetically related obligate parasites.

Conclusion

The functions of protist polyketides are unknown, but since they are structurally different from those of bacterial and fungal origin it is not unreasonable to expect they may have diverse and divergent functions. In any case, the diversity of modular Type I PKS products seems to arise from frequent recombination events among the modules. The modular PKS system provides an extraordinary platform for recombination, with the evolutionary advantage that the organisms have the ability to produce a large chemical diversity from a limited number of different genes.

Novel polyketides often exhibit bioactivity, with potentially valuable functions as antibiotics against bacterial, fungal or protozoan pathogens. The protistan lineages have not been extensively investigated for novel polyketides (with exception of the known potent polyether biotoxins) and therefore may represent a vast pool of undiscovered bioactive substances, particularly critical at a time when the severe clinical problem of multidrug resistant pathogens is on the rise.

Recent inferences on biosynthetic pathways and secondary metabolite structures of bioactive compounds based on comparative gene sequence information in cyanobacteria (Sudek et al. 2006) points the way to similar future discoveries from protists. Future work is aimed at elucidating the structure of protistan polyketides and determining the genetic regulation of transcription of PKS genes to clarify the functional significance of these secondary metabolites.

Methods

Data sets: Our base data set was the alignment from Kroken et al. (2003) containing a representative subset of KS domains from bacterial and fungal PKS, metazoan FAS and from oxoacyl-ACP

synthases. Newly discovered protistan KS sequences were added to a random selection of sequences from each large clade from this data set and completely re-aligned using different methods (see Table 1 and Supplementary Table S1).

Screening: The above data sets, the public databases available on the NCBI website and eukaryotic genome sequences from COGENT (Janssen et al. 2003) were screened for sequences showing similarities to the already known protistan PKS from *C. parvum* (Genbank accession EAK87820) using protein—protein (BLASTp) and protein—-nucleotide (tBLASTn) BLAST (Altschul et al. 1997). We identified genomic regions containing multiple fragments with similarity to different PKS domains as putative PKS genes. KS domain sequences longer than 300 amino acids, and showing significant similarity ($e < 10^{-5}$) were used in the multiple alignments and phylogenetic analysis.

Tentative assembly of *Emiliania huxleyi* PKS gene candidates: At the time of our analyses, no assembly for the *E. huxleyi* genome was available. Thus, we have assembled genomic segments containing putative PKS fragments as follows. The sequence and quality files of the 3,868,934 genomic reads available (as of October 2005) from the NCBI trace database were downloaded and screened with tBLASTn with the complete amino acid sequence of the *C. parvum* PKS (EAK87820) as query. All highly significant hits were re-blasted (BLASTn) against the traces to identify potentially overlapping reads. This selection of reads was assembled using phrap (http://www.phrap.org/) into 389 contigs, of which 17 were longer than 10,000 bases. Six of these contigs contained apparent frame-shifts. From the remaining eleven (containing combinations of all known PKS domains in large contiguous ORFs), five KS domains chosen from two contigs (showing the highest similarity to other protistan KS sequences) were further analyzed.

After the recent publication of the first draft of the *E. huxleyi* assembly, we checked for the presence of the KS domain sequences in the public assembly, which was cleaned of contaminating nucleotide sequences. The origin of these sequences from *E. huxleyi* was thus confirmed.

Alignments and phylogenetic analyses: Simultaneous HMM-based alignment and tree construction were performed with SATCHMO (Edgar and Sjölander 2003). SatchmoView was used to inspect the results and identify well-conserved regions for alignment within subgroups of the sequences. Profile hidden Markov models (HMM) for these regions were calculated and calibrated using hmmbuild and hmmcalibrate from the HMMer package (http://hmmer.wustl.edu/). HMM logos for these models were generated from the HMM Logo web server "LogoMat-M" (http://logos.molgen.mpg.de/) (Schuster-Böckler et al. 2004). HMM searches with the profile models of regions of group-specific or overall conservation were performed with hmmsearch from the HMMer package. For phylogenetic analyses, multiple alignments were prepared using kalign (Lassmann and Sonnhammer 2005). The alignment contained 130 sequences and 679 characters (provided upon request). Maximum likelihood phylogenetic trees were calculated with PhyML (Guindon and Gascuel 2003) using a BIO-NJ tree as starting tree, the WAG evolutionary model, with a gamma distribution parameter estimated from the data. Bootstrap analyses were performed with the same settings in 200 replicates. Bayesian analyses were conducted via MrBayes v.3.1.2 (Ronquist and Huelsenbeck 2003), with one million iterations, two runs with six chains each, a temperature parameter of 0.05 and with the WAG amino acid model. Burnin for construction of the consensus tree was 500,000 generations. The program Asatura (van de Peer et al. 2002) was used to explore substitution saturation in the multiple alignments.

Acknowledgements

Arthur Grossmann (Stanford University, USA) and Brian Palenik (UC San Diego, USA) kindly provided the *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus* PKS sequences, respectively. Christian Hertweck (HKI, Germany) contributed to fruitful discussions. This research was partly funded by the EU projects EUKETIDES (QLK3-CT-2002-01940), ESTTAL (GOCE-CT2004-511154), and Network of Excellence (NoE) Marine Genomics Europe (GOCE-CT-2004-505403).

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.protis.2007.08.001.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**: 3389—3402

Baldauf SL (2003) The deep roots of eukaryotes. Science **300**: 1703—1706

Bhattacharya D, Medlin L (1998) Algal phylogeny and the origin of land plants. Plant Physiol **116**: 9—15

Cembella AD, John U (2006) Molecular Physiology of Toxin Production and Growth Regulation in Harmful Algae. In Graneli E, Turner JT (eds) Ecology of Harmful Algae. Springer-Verlag, Heidelberg, pp 215—227

Derelle E, Ferraz C, Lagoda P, Eychenie S, Cooke R, Regad F, Sabau X, Courties C, Delseny M, Demaille J, Picard A, Moreau H (2002) DNA libraries for sequencing the genome of *Ostreococcus tauri* (Chlorophyta, Prasinophyceae): The smallest free-living eukaryotic cell. J Phycol **38**: 1150—1156

Derelle E, Ferraz C, Rombaut S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piégu B, Ball SG, Ral JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA **103**: 11647—11652

Edgar RC, Sjölander K (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics **19**: 1404—1411

Gross F, Luniak N, Perlova O, Gaitatzis N, Jenke-Kodama H, Gerth K, Gottschalk D, Dittmann E, Muller R (2006) Bacterial type III polyketide synthases: phylogenetic analysis and potential for the production of novel secondary

metabolites by heterologous expression in pseudomonads. Arch Microbiol **185**: 28—38

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52**: 696—704

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

Jaeckisch N, Singh R, Curtis C, Cembella AD, John U (2007). Genomic Characterization of the Spirolide-Producing Dinoflagellate *Alexandrium ostenfeldii* with Special Emphasis on PKS Genes. In Moestrup Ø et al. (eds) XIIth International Conference on Harmful Algae (2006). Intergovernmental Oceanographic Commission of UNESCO, Paris, in press

Janssen PJ, Enright AJ, Audit B, Cases I, Goldovsky L, Harte N, Kunin V, Ouzounis CA (2003) COGENT: a flexible data environment for computational genomics. Bioinformatics **19**: 1451—1452

Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol **22**: 2027—2039

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW (2006) The tree of eukaryotes. Trends Ecol Evol **20**: 670—676

Keeling PJ, Slamovits CH (2005) Causes and effects of nuclear genome reduction. Curr Opin Genet Dev **15**: 601—608

Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc Natl Acad Sci USA **100**: 15670—15675

Lassmann T, Sonnhammer ELL (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics **6**: 298

Lidie KB, Ryan JC, Barbier M, Van Dolah FM (2005) Gene expression in Florida red tide dinoflagellate *Karenia brevis*: analysis of an expressed sequence tag library and development of DNA microarray. Mar Biotech **7**: 481—493

MacKinnon SL, Cembella AD, Burton IW, Lewis N, LeBlanc P, Walter JA (2006) Biosynthesis of 13-Desmethyl Spirolide C by the dinoflagellate *Alexandrium ostenfeldii*. J Org Chem **71**: 8724—8731

Moore BS, Hopke JN (2001) Discovery of a new bacterial polyketide pathway. Chem Bio Chem **2**: 35—38

Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otillar R, Merchant SS, Podell S, Gaasterland T, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek S, Heijde M, Jabbari K, Bowler C, Lohr M, Robbens S, Werner G, Dubchak I, Pazour GJ, Ren Q, Paulsen I, Delwiche C, Schmutz J, Rokhsar D, Van de Peer Y, Moreau H, Grigoriev I (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci USA **104**: 7705—7710

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**: 1572—1574

Schuster-Böckler B, Schultz J, Rahmann S (2004) HMM Logos for visualization of protein families. BMC Bioinformatics **5**: 1—8

Shen B, Du L, Sanchez C, Edwards DJ, Chen M, Murrel JM (2001) The biosynthetic gene cluster for the anticancer drug bleomycin from *Streptomyces verticillus* ATCC15003 as a model for hybrid peptide—polyketide natural product biosynthesis. J Ind Microbiol Biotechnol **27**: 12961—12964

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

Snyder RV, Guerrero MA, Sinigalliano CD, Winshell J, Perez R, Lopez JV, Rein KS (2005) Localization of polyketide synthase encoding genes to the toxic dinoflagellate *Karenia brevis*. Phytochemistry **66**: 1767—1780

Staunton J, Weissman KJ (2001) Polyketide biosynthesis: A millennium review. Nat Prod Rep **18**: 380—416

Sudek S, Haygood MG, Youssef DTA, Schmidt EW (2006) Structure of trichamide, a cyclic peptide from the bloom-forming cyanobacterium *Trichodesmium erythraeum*, predicted from the genome sequence. Appl Environ Microbiol **72**: 4382—4387

Van de Peer Y, Frickey T, Taylor JS, Meyer A (2002) Dealing with saturation at the amino acid level: a case study involving anciently duplicated zebrafish genes. Gene **295**: 205—211

Wright JLC, Cembella AD (1998) Ecophysiology and Biosynthesis of Polyether Marine Biotoxins. In Anderson DM, Cembella AD, Hallegraeff GM (eds) Physiological Ecology of Harmful Algal Blooms. Springer-Verlag, Heidelberg, pp 427—451

Zhu G, Li Y, Cai X, Millership JJ, Marchewka MJ, Keithly JS (2004) Expression and functional characterization of a giant Type I fatty acid synthase (CpFAS1) gene from *Cryptosporidium parvum*. Mol Biochem Parasitol **134**: 127—135

Zhu G, LaGier MJ, Stejskal F, Millership JJ, Cai X, Keithly JS (2002) *Cryptosporidium parvum*: The first protist known to encode a putative polyketide synthase. Gene **298**: 79—89