Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus*

Ann Pearson*†, Meytal Budin*, and Jochen J. Brocks‡

Departments of *Earth and Planetary Sciences and ‡Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Communicated by Andrew H. Knoll, Harvard University, Cambridge, MA, October 10, 2003 (received for review July 9, 2003)

Sterol biosynthesis is viewed primarily as a eukaryotic process, and the frequency of its occurrence in bacteria has long been a subject of controversy. Two enzymes, squalene monooxygenase and oxidosqualene cyclase, are the minimum necessary for initial biosynthesis of sterols from squalene. In this work, 19 protein gene sequences for eukaryotic squalene monooxygenase and 12 protein gene sequences for eukaryotic oxidosqualene cyclase were compared with all available complete and partial prokaryotic genomes. The only unequivocal matches for a sterol biosynthetic pathway were in the proteobacterium, *Methylococcus capsulatus*, in which sterol biosynthesis is known, and in the planctomycete, *Gemmata obscuriglobus*. The latter species contains the most abbreviated sterol pathway yet identified in any organism. Analysis shows that the major sterols in *Gemmata* are lanosterol and its uncommon isomer, parkeol. There are no subsequent modifications of these products. In bacteria, the sterol biosynthesis genes occupy a contiguous coding region and possibly comprise a single operon. Phylogenetic trees constructed for both enzymes show that the sterol pathway in bacteria and eukaryotes has a common ancestry. It is likely that this contiguous reading frame was exchanged between bacteria and early eukaryotes via lateral gene transfer or endosymbiotic events. The primitive sterols produced by *Gemmata* suggest that this genus could retain the most ancient remnants of the sterol biosynthetic pathway.

Sterol biosynthesis is nearly ubiquitous among eukaryotes; conversely, it is almost completely absent in prokaryotes (1). As a result, the presence of diverse steranes in ancient rocks is used as evidence for eukaryotic evolution >2.7 billion years ago (2). However, the occasional presence of sterols in prokaryotes is poorly understood. Sterol production by bacteria previously has been demonstrated only in the Methylococcales (3, 4) and Myxobacteriales (5, 6).

Understanding the evolution of sterol biosynthesis is of significant interest to biochemistry, evolutionary biology, and the geosciences, because the only known biosynthetic pathway requires molecular oxygen. The first step in this pathway is the epoxidation of the hydrocarbon squalene, in which the addition of $\frac{1}{2}O_2$ is catalyzed by the enzyme squalene monooxygenase (SQMO) (7). Unless there are other unknown enzymes or abiogenic reactions capable of producing squalene epoxide, this would require the prior evolution of oxygenic photosynthesis. For sterol biosynthesis to date to the last common ancestor, a biogenic or abiogenic peroxidation could be a potential mechanism, although this has not yet been demonstrated.

Cyclization of squalene epoxide to form the initial sterol proceeds immediately through the action of a second enzyme, oxidosqualene cyclase (OSC). It is believed that OSC evolved from the hopanoid pathway predecessor, bacterial squalene–hopene cyclase (SHC) (8, 9). In eukaryotes, the initial sterols lanosterol and cycloartenol are merely biosynthetic intermediates; i.e., up to 20 additional steps are required for downstream conversion of lanosterol to the animal product, cholesterol. Synthesis of sterols, however, requires only SQMO and OSC. The major outstanding questions are when and in what organisms did the original synthesis pathway develop?

Here we conducted an exhaustive search of all microbial genetic sequence data currently in the public domain to (*i*) identify the presence of SQMO genes in prokaryotes, (*ii*) identify the OSC genes carried by organisms also containing SQMO and determine whether these sequences are consistent with lanosterol synthase, cycloartenol synthase, or SHC activity, and (*iii*) investigate the phylogeny of these genetic sequences with respect to sterol biosynthesis in eukaryotes. We hypothesized that sequence similarity to a specific group of eukaryotes (animals, plants, fungi, or protists) would be consistent with a recent lateral gene transfer event, whereas divergence of the bacterial sequences into a basal clade would indicate an ancient biochemical origin. Significantly, these results could impact our understanding of the coevolution of metabolic pathways, the composition of the atmosphere, and the radiation of nucleated cells (eukaryotes).

Methods

Genome Searches. Amino acid sequences for all identified and annotated eukaryotic SQMO genes were compiled from the National Center for Biotechnology Information (NCBI) and Department of Energy (DOE)/Joint Genome Institute (JGI) public databases, www.ncbi.nih.gov and www.jgi.doe.gov (see the first column of Table 2, which is published as supporting information on the PNAS web site). Amino acid sequences for OSC genes of the same organisms also were obtained when available (Table 2, second column). To identify SQMO genes in prokaryotes, each SQMO sequence from Table 2 was searched by BLAST (using default values) against all complete and partial microbial genomes available through NCBI, as well as the genomes currently in process at the Department of Energy Joint Genome Institute and the Max Planck Institute (http://blast.mpi-bremen.de/blast; accessed March, 2003). The best matches (Table 1) were defined as all species that yielded at least one expect similarity value $<10^{-8}$ and were compiled and ranked by weighted similarity score as described below. Because many microbial species contain OSC genes, including SHCs, searches for OSC genes were conducted only for those organisms identified through the SQMO similarity searches. Phylogeny of the SHCs is not discussed in this work.

Raw DNA sequence data for the putative SQMO and OSC coding regions of *M. capsulatus* (gnl TIGR_414 contig: 229:Methylococcus capsulatus) and *G. obscuriglobus* (gnl TIGR_214688 contig:782:Gemmata obscuriglobus UQM 2246) were translated into all six reading frames with minimum read length set to 300 aa. Resulting translated sequences were BLASTed against the Swiss-Prot protein sequence database to identify putative functional homologues.

Sequence Alignments. All data were aligned as protein sequences. Multiple alignments were done in CLUSTALW 1.8, using the BLOSUM weight matrix, gap open penalty 10, gap extension penalty 0.0, and

Abbreviations: OSC, oxidosqualene cyclase; SHC, squalene–hopene cyclase; SQMO, squalene monooxygenase.

†To whom correspondence should be addressed. E-mail: pearson@eps.harvard.edu.

www.pnas.org/cgi/doi/10.1073/pnas.2536559100

Table 1. BLAST results for SQMO and OSC genes of eukaryotes (rows) vs. prokaryotes (columns)

	Methylococcus capsulatus	*Gemmata obscuriglobus*	*Burkholderia mallei*	*Bradyrhizobium japonicum*	*Azotobacter vinelandii*	*Nostoc punctiforme**	*Pseudomonas aeruginosa**	*Bacillus halodurans**	*Thermobifida fusca*	*Coxiella burnetii*
Arabidopsis† *thaliana* (ER11)	—/1e^{-100}	7e^{-7}/3e^{-80}	3e^{-06}/2e^{-45}	—/6e^{-37}	—/2e^{-43}	—/1e^{-41}	—/—	3e^{-06}/—	—/—	—/—
A. thaliana (ER12)	—	2e^{-6}	7e^{-05}	—	—	—	—	5e^{-05}	—	—
A. thaliana (ER13)	2e^{-15}	9e^{-13}	2e^{-12}	2e^{-7}	4e^{-4}	1e^{-13}	9e^{-07}	4e^{-10}	4e^{-5}	4e^{-10}
A. thaliana (NP564)	1e^{-13}	1e^{-14}	5e^{-08}	2e^{-5}	4e^{-3}	1e^{-07}	2e^{-07}	1e^{-06}	1e^{-5}	—
A. thaliana (NP568)	2e^{-19}	2e^{-13}	3e^{-12}	2e^{-4}	1e^{-3}	8e^{-03}	9e^{-08}	—	7e^{-5}	—
Brassica napus (ER11)	1e^{-10}/n.a.	1e^{-07}/n.a.	3e^{-07}/n.a.	—/n.a.	—/n.a.	—/n.a.	1e^{-03}/n.a.	1e^{-06}/n.a.	—/n.a.	—/n.a.
B. napus (ER12)	5e^{-7}	8e^{-06}	4e^{-05}	—	—	—	—	3e^{-05}	—	—
Candida albicans	2e^{-12}/2e^{-87}	1e^{-18}/4e^{-78}	5e^{-03}/3e^{-29}	3e^{-07}/8e^{-15}	8e^{-05}/9e^{-25}	9e^{-05}/1e^{-25}	6e^{-08}/—	3e^{-02}/—	5e^{-3}/—	—/—
Homo sapiens	7e^{-16}/1e^{-120}	8e^{-15}/6e^{-92}	1e^{-04}/5e^{-50}	4e^{-02}/2e^{-47}	—/4e^{-38}	2e^{-05}/2e^{-37}	4e^{-07}/—	1e^{-07}/—	—/—	—/—
Leishmania major	6e^{-14}/n.a.	6e^{-13}/n.a.	—/n.a.	—/n.a.	—/n.a.	—/n.a.	—/n.a.	—/n.a.	—/n.a.	—/n.a.
Medicago truncatula (49)	2e^{-20}/1e^{-60}	6e^{-19}/2e^{-52}	1e^{-15}/1e^{-36}	9e^{-10}/9e^{-30}	1e^{-10}/4e^{-32}	9e^{-05}/6e^{-28}	3e^{-17}/—	1e^{-02}/—	1e^{-7}/—	1e^{-8}/—
M. truncatula (48)	3e^{-14}	3e^{-15}	3e^{-10}	5e^{-3}	2e^{-4}	4e^{-02}	2e^{-09}	—	3e^{-2}	—
Mus musculus	9e^{-16}/1e^{-118}	3e^{-15}/2e^{-92}	3e^{-09}/5e^{-52}	2e^{-5}/7e^{-46}	4e^{-3}/8e^{-37}	3e^{-07}/2e^{-35}	2e^{-08}/—	1e^{-09}/—	3e^{-3}/—	—/—
Oryza sativa (687)	7e^{-17}/1e^{-101}	2e^{-14}/6e^{-79}	3e^{-10}/2e^{-39}	7e^{-6}/5e^{-41}	3e^{-6}/2e^{-39}	4e^{-02}/3e^{-35}	6e^{-08}/—	7e^{-07}/—	4e^{-6}/—	—/—
O. sativa (686)	1e^{-18}	2e^{-13}	2e^{-08}	1e^{-4}	5e^{-6}	—	5e^{-08}	—	1e^{-3}	—
Panax ginseng	2e^{-18}/1e^{-103}	3e^{-18}/2e^{-83}	1e^{-13}/9e^{-41}	3e^{-6}/5e^{-36}	3e^{-7}/3e^{-38}	6e^{-06}/2e^{-39}	9e^{-10}/—	—/—	6e^{-10}/—	6e^{-10}/—
Rattus norvegicus	1e^{-14}/1e^{-119}	2e^{-16}/1e^{-93}	5e^{-10}/1e^{-54}	6e^{-4}/2e^{-46}	4e^{-4}/5e^{-39}	7e^{-08}/9e^{-38}	2e^{-09}/—	2e^{-11}/—	7e^{-3}/—	5e^{-4}/—
Saccharomyces cerevisiae	2e^{-11}/9e^{-86}	6e^{-18}/5e^{-75}	4e^{-2}/1e^{-17}	1e^{-4}/1e^{-19}	5e^{-2}/2e^{-21}	1e^{-03}/3e^{-24}	8e^{-08}/—	4e^{-02}/—	6e^{-2}/—	—/—
Schizosaccharomyces pombe	1e^{-13}/6e^{-98}	1e^{-15}/1e^{-82}	3e^{-5}/2e^{-39}	3e^{-4}/4e^{-35}	—/8e^{-34}	—/1e^{-33}	6e^{-04}/—	1e^{-04}/—	—/—	—/—
Weighted score	(13|82)	(13|67)	(7|33)	(3|29)	(3|28)	(3|28)	(6|0)	(4|0)	(3|0)	(2|0)

Expect values are presented in the order SQMO/OSC. Single copies of OSC genes are scored in the first row for each organism; multiple copies of SQMO genes are scored separately. The weighted average probability scores are shown in the last row. n.a., sequence not available; —, no hits.

*Complete genome.

†Accession numbers and sequence information are given in Table 2.

hydrophilic gaps. The output was checked manually for alignment of expected motifs before proceeding to phylogenetic tree building. Length heterogeneity at the beginning and ends of sequences were masked, but no indel gaps were removed.

Phylogenetic Tree Construction. Unrooted phylogenetic trees were created by using PHYLIP 3.6α3 (http://evolution.genetics.washington.edu/phylip/phylip36.html). Distance matrices were calculated by using the PROTDIST function, with the Jones–Taylor–Thornton matrix and equal character weighting. The output was used to create neighbor-joining trees in the NEIGHBOR distance analysis program, with randomized order of sequence input and negative branch lengths allowed. Parsimony analysis was performed by using the PROTPARS function, with randomized sequence addition (jumble 10 times). Gaps were treated as "missing." Maximum-likelihood analysis was performed by using PROML, also with the Jones–Taylor–Thornton matrix, unweighted characters, and randomized order of input (jumble 10 times). For each analysis, a total of 100 bootstrap analyses were performed, and global rearrangements were applied when available. Consensus trees were calculated from the raw output, and bootstrap confidence is presented in the order neighbor-joining/parsimony/maximum likelihood.

Subsequent realignment and neighbor-joining calculations applied only to the active FAD binding regions of the SQMO genes did not generate significantly different conclusions and thus are not discussed.

Cultures. *Gemmata* strain DSMZ 5831 (Deutsche Sammlung von Mikroorganismen und Zellkulturen) was purchased as desiccated cells and subsequently rehydrated in PYGV medium (0.25 g of peptone/0.25 g of yeast extract/0.25 g of dextrose/1.0 liter of H_2O). *Gemmata* strain Wa1-1 was obtained in live culture from the University of Washington, Seattle. Both DSMZ 5831 and Wa1-1 cells were started on PYGV agar plates, then transferred to liquid medium. Liquid medium was prepared at 10× concentration, extracted three times with methylene chloride and hexane, diluted to 1 liter, and autoclaved for 25 min. Cells for lipid analysis were grown in PYGV containing 0.1% cycloheximide, a fungal inhibitor. An identical, cell-free negative control was maintained and subsequently processed for lipid analysis.

Lipid Analysis. Initial samples of centrifuged cells were extracted with CH_2Cl_2 to obtain total extractable lipids. Subsequent samples were extracted successively: cells were suspended in H_2O and extracted five times with CH_2Cl_2, with ultrasonication and vortexing. Cellular debris was then captured on a combusted quartz fiber filter; half of the filter was hydrolyzed in KOH/CH_3OH (90°C for 1 h) and half was hydrolyzed in H_2O/HCl (70°C for 4 h). All samples were dried over Na_2SO_4 and derivatized with bis(trimethylsilyl)trifluoroacetamide (5% trimethylchlorosilane) in pyridine. Gas chromatography/MS analyses were performed on an Agilent Technologies (Palo Alto, CA) 6890 gas chromatograph coupled to a 5973 mass selective detector, equipped with a 60-m CP-Sil5 column (1% phenyl). Free lanosterol and parkeol were identified by their mass spectra, coinjection of authentic standard (lanosterol), 500-MHz ^{1}H NMR [lanosterol and parkeol; purified by using an Agilent Technologies 1100 LC-MS with an atmospheric pressure chemical ionization (APCI) source], and reconfirmation of coelution on a second gas chromatography column of different polarity [30-m DB-5 column (5% phenyl)].

Results and Discussion

Genome Searches. Our comparison of all bacterial and archaeal complete and partial genomes in the public domain with the SQMO and OSC amino acid sequences from eukaryotes yielded only two significant matches: *M. capsulatus* and the planctomycete, *G. obscuriglobus*. There were no significant matches among the archaea. The bacterial species showing significant similarity to eukaryotic SQMO genes are presented in columnar order from most to least significant expect value (Table 1). A weighting method was devised to determine an overall probability score for

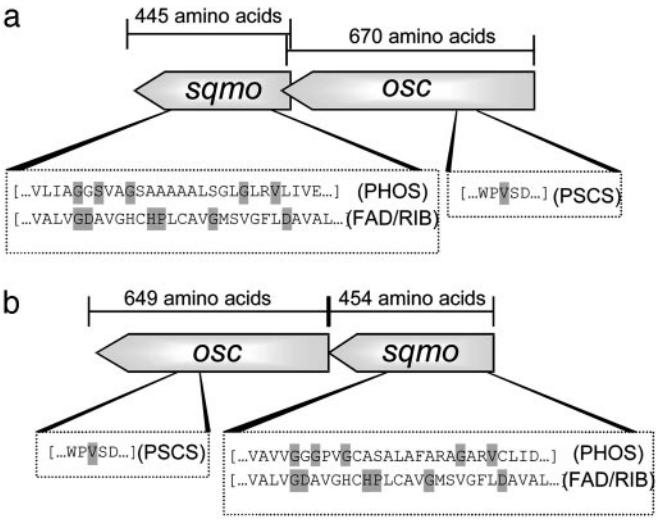

Fig. 1. The sterol biosynthesis genes found in *M. capsulatus* (*a*) and *G. obscuriglobus* (*b*), showing the direction of transcription (arrows) and the critical conserved amino acids (expanded view; critical residues are highlighted). Important functional regions (25) PSCS (protosteryl cation stabilizing), PHOS (pyrophosphate binding), and FAD/RIB (FAD-ribityl binding) are discussed in the text.

each species in the table. It was calculated as the absolute value of the average exponent of the expect values, or

$$\text{weighted score} = \sum[-\log(E)]/N,$$

where E is the expect value and N is the total number of eukaryotic sequences searched. The bacterium with the highest weighted probability for sterol biosynthesis is *M. capsulatus*, a species already known to produce sterols. The surprising result was the nearly equal score set obtained for *Gemmata*. Translation of the raw DNA sequences containing these regions of high similarity showed that the genes are contiguous in both species and may be part of a single operon (Fig. 1). At the present time, *G. obscuriglobus* and *M. capsulatus* represent the only prokaryotic species found to have the definitive genetic capacity for sterol synthesis. However, further sequencing of myxobacteria is required. There are no sequences available for the myxobacteria, *Nannocystis excedens* and *Stigmatella aurantiaca*, but the recent cloning of an OSC gene from *S. aurantiaca* (5) suggests that SQMO is likely to be found if appropriate sequencing efforts are initiated.

The only other bacterium yielding data of numerical significance across the majority of eukaryotic SQMO and OSC sequences was *B. mallei*. Comparison of the target proteins against the Swiss-Prot database revealed the cause of the similarity. In these species, the OSC searches are detecting a putative SHC, whereas the SQMO searches are detecting a homologous FAD-binding monooxygenase with similarity to the UbiH or UbiF ubiquinone monooxygenases. The genes do not have adjacent loci, a finding contrary to what would be expected for SQMO and OSC. Further experiments should be done to determine whether sterols indeed are absent in the *Burkholderia*.

The absence of SQMO genes in other bacterial groups was not surprising; most reports of sterols in bacteria subsequently have been attributed to analytical contamination or uptake from exogenous sources (1). Examples include cultures of cyanobacteria (10) and mycobacteria (11). In the former case (10), it is likely that the detected sterols resulted from contamination of the cultures by rust fungus.[§] In the latter case (11), the medium contained yeast extract, potentially a rich source of exogenous sterols. The validation assay in the latter work (^{14}C-mevalonate incorporation) also was specific for the MVA pathway of isoprenoid biosynthesis, which the mycobacteria do not possess (12). We surmise that the ^{14}C-mevalonate was metabolized and incorporated into other cellular lipids via the methylerythritol phosphate (MEP) pathway (13), leading to the erroneous conclusion that mycobacteria produce sterols. No SQMO gene homologues were found in the genomes of the mycobacteria (including whole genomes of *Mycobacterium bovis*, *Mycobacterium leprae*, and *Mycobacterium tuberculosis*) or in the cyanobacteria (including whole genomes of the genera *Nostoc*, *Prochlorococcus*, *Synechococcus*, *Synechocystis*, *Thermosynechococcus*, and *Trichodesmium*).

***Gemmata* Cultures.** The discovery of sterol biosynthesis genes in *Gemmata* was especially intriguing. *Planctomycetales* are budding eubacteria that lack peptidoglycan in their cell walls and contain intracellular membranes (14, 15). Unusual lipid biosynthetic capabilities recently were demonstrated in the "anammox" group (16) of planctomycetes: *Candidatus B. anammoxidans* and *Candidatus K. stuttgartiensis* synthesize ladderanes (17), which are thought to be significant for intracellular partitioning. Although aspects of cellular organization vary among genera, *Gemmata* have attracted particular attention because of their morphological resemblance to unicellular eukaryotes. Uniquely among all bacteria, *Gemmata* have a double bilayer nuclear membrane (18). Recently, an analysis of 16S rRNA phylogeny based on only the slowly evolving regions of this molecule suggested that the *Planctomycetales* may be a deeply branching bacterial clade (19), though this result is still controversial because many *Planctomycetales* are neither thermophilic nor anaerobic, the presumed physiology of the last common ancestor.

We cultured two strains of *Gemmata* to test for the presence of sterols, *G. obscuriglobus* strain DSMZ 5831 and *Gemmata* strain Wa1-1 (an unnamed species of the genus *Gemmata* with 93% 16S rRNA similarity). In accordance with recent recommendations,[§] single colonies were harvested and inoculated into PYGV medium containing 1 μl/ml cycloheximide, a fungal growth inhibitor. Preliminary work showed that failure to include cycloheximide resulted in trace contamination by a suite of sterols similar to what is found in cyanobacterial cultures contaminated with rust fungus. All media were preextracted with solvents to remove any exogenous lipids, including sterols and sterol precursors. Comprehensive process blanks yielded no detectable sterols.

Our experiments show that *Gemmata* cells contain the simplest suite of sterol products yet found in any organism. Analysis of the sterol composition of *Gemmata* first was attempted for a total lipid extract of whole cells. Very minor quantities of lanosterol [lanosta-8(9)-3β-ol] and parkeol [lanosta-9(11)-3β-ol] were detected. Subsequent saponification of the polar lipids did not release any additional sterols. Therefore, the extractable lipid fractions of these cultures essentially were devoid of sterols, explaining why sterols in *Gemmata* formerly have gone unnoticed.

However, when the cellular residue from the previous extractions was hydrolyzed directly, either in H_2O/HCl or by saponification, large amounts of sterol were released. The sterols of *Gemmata* exclusively are the C_{30} compounds, lanosterol and parkeol (Fig. 2). No $C_{27\text{-}29}$ sterols were detected in the samples, and there was no significant difference between strains Wa1-1 and DSMZ 5831. Calculations of absolute and relative amounts show that sterols are present in approximately a 1:1 mole ratio to the major fatty acid component, steric acid ($C_{18:0}$), and at a total of nearly 20 mg/g cell. The minor amount of free sterol found initially was inconsequential when compared with the bound fraction. These high concentrations exceed any previous report of sterols in prokaryotes. *M. capsulatus* contains 2.2–3.5 mg/g total sterols (3, 20, 21), and 4 mg/g was reported in *N. excedens* (6). The hydrolysis-labile extraction suggests that the sterols are bound to macromolecular constituents of the

[§]Summons, R. E., Jahnke, L. L., Cullings, K. W. & Logan, G. A. (2001) *Trans. Am. Geophys. Union* **82,** Fall Meeting Suppl., Abstr. B22D-0184.

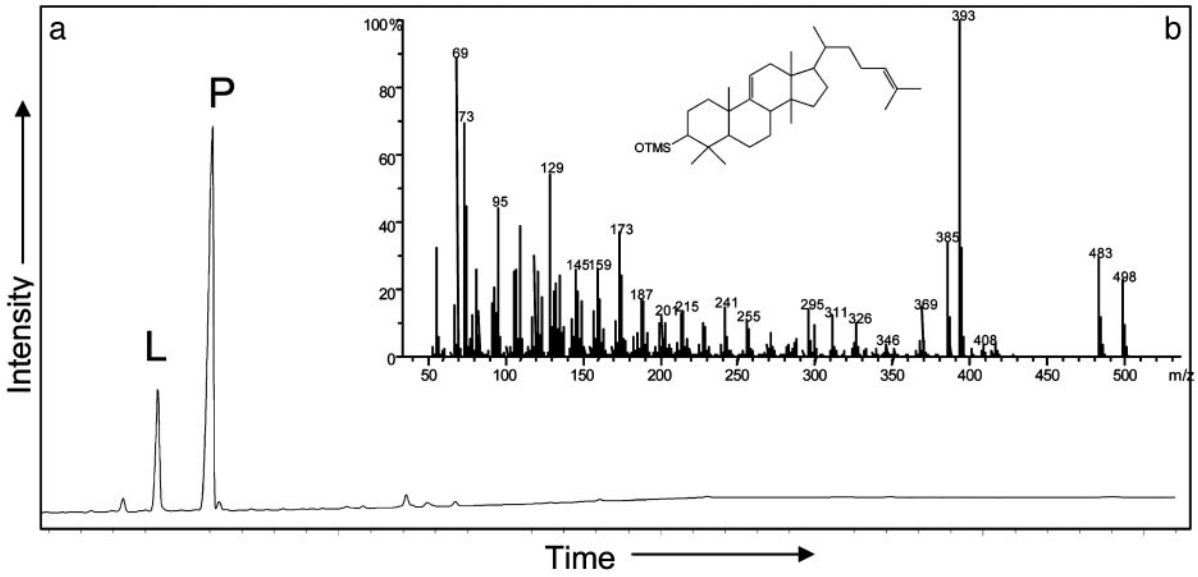

Fig. 2. (*a*) Chromatogram of the sterol region of acid-hydrolyzed whole cells, showing lanosterol (L) and parkeol (P). Minor peaks are nonsteroidal, with the possible exception of the peak immediately to the right of parkeol. (*b*) Structure and mass spectrum (70 eV; 1 eV = 1.602×10^{-19} J) of parkeol as its trimethylsilyl derivative. See Figs. 6 and 7 and Table 3, which are published as supporting information on the PNAS web site.

insoluble cell material and could be associated with the cell walls. In contrast, sterols in *M. capsulatus* and *N. excedens* are extractable as free components.

During sterol biosynthesis in eukaryotes, the initial C_{30} biosynthetic products primarily are lanosterol and cycloartenol. Extensive modifications lead to end-products such as cholesterol (C_{27}; animals), ergosterol (C_{28}; fungi), and sitosterol (C_{29}; plants). The first steps in these complicated rearrangements involve demethylation, initially at position C-14 and subsequently at C-4. In *Gemmata*, the lack of products $<C_{30}$ indicates that the synthesis stops short of demethylation at C-14; indeed, a 14α-demethylase has not yet been identified in *Gemmata*. This is not the case for the *Methylococcales*, in which modification to 4,4-dimethyl (C_{29}) and 4-methyl (C_{28}) products is observed (3, 4, 20, 21), while the myxobacteria (5, 6) appear to make products throughout the range C_{27}–C_{30}. The absence of any products other than C_{30} compounds in *Gemmata* indicates that this group never evolved or acquired any of the enzymes downstream from OSC.

Synthesis of the initial C_{30} isomers is controlled by the carbocation intermediate formed during cyclization of squalene epoxide by OSC (22). Different forms of OSC enzymes produce different isomers. All plants contain cycloartenol synthase and produce only cycloartenol; similarly, all animals and fungi contain lanosterol synthase and produce only lanosterol. The production of a third isomer, parkeol, occurs only under exceptional circumstances. In each case, formation of the carbocation intermediate is governed by the protosteryl cation stabilizing (PSCS) region (Fig. 1). The PSCS of all cycloartenol synthases contains the amino acid motif "WPI" (Fig. 3). The isoleucine (I) is responsible for the exclusive production of cycloartenol. In the same location, all lanosterol synthases contain either "WIV" or "YTV" (Fig. 3). The valine (V) is responsible for the production of lanosterol. However, laboratory-induced I $\rightarrow$ V mutants of cycloartenol synthases ("WPV") form cycloartenol, lanosterol, and parkeol (23). This phenomenon is known as nondiscriminate cyclization. The WPV in *Gemmata* (and *M. capsulatus*) is a natural occurrence of a nondiscriminate OSC. It is unknown why *Gemmata* produces parkeol in excess of lanosterol or why cycloartenol is absent, but other regions of the enzyme evidently also contribute to the final isomeric ratio of products. We hypothesize that the WPV sequence found in bacteria reflects an ancient form of the OSC gene. There may be no evolutionary pressure toward production of a single isomer, because there are no downstream enzymatic modifications that would require a specific substrate.

Similar patterns of divergence can be found among the plant, animal/fungal, and bacterial groupings of SQMO sequences (Fig. 4). Positions 126–153 correspond to the GxGxxG motif that binds the pyrophosphate groups of FAD (24). This pattern is preserved in all species examined. In contrast, there is variation in the region that binds the ribityl group of FAD, corresponding to positions 401–431 (24). Distinct patterns are apparent and again bacteria show characteristics intermediate between the forms possessed by plants and animals/fungi. All animals and fungi contain the motif "LTGG," whereas plants are characterized by "xxAS" in the same position. *M. capsulatus* and *G. obscuriglobus* fall into neither pattern, and each shows elements of both eukaryotic motifs: "LTAS" in *M. capsulatus* and "LCAV" in *Gemmata*. These unique sequences again can be used to suggest that the bacterial genes represent an evolutionarily distinct lineage. It would not be difficult to imagine the divergence of the two eukaryotic motifs from a precursor of the form LTAS.

```
                                [477       487]
CANDIDA_ERG7                  ...QGYTVSDCTAE...
SACCHAROMYCES_ERG7            ...QGYTVADCTAE...
SCHIZOSACCHAROMYCES_OSC       ...QGYTVSDTTSE...
MUS_ERG7                      ...CGWIVADCTAE...
RAT_ERG7                      ...CGWIVADCTAE...
HOMO_ERG7                     ...CGWIVSDCTAE...
ORYZA_CAS                     ...HGWPISDCTAE...
PANAX_CAS                     ...HGWPISDCTAE...
ARABADOPSIS_CAS1              ...HGWPISDCTAE...
STIGMATELLA_OSC               ...HGWPISDCTAE...
METHYLOCOCCUS_OSC             ...HCWPVSDCAAE...
GEMMATA_OSC                   ...HAWPVSDCTAE...
                                  : ::* ::*
```

Fig. 3. Protein sequence alignment showing the protosteryl cation-binding functional region of OSCs. Numbering corresponds to the amino acid position numbers of *Arabidopsis* CAS1.

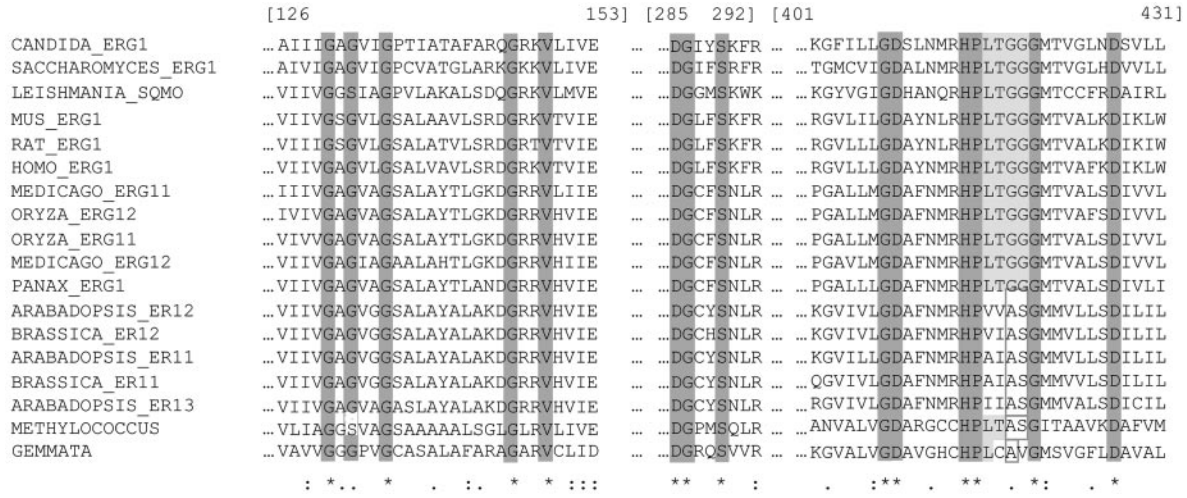

Fig. 4. Protein sequence alignment showing the FAD-binding functional regions of SQMOs. Numbering corresponds to the amino acid position numbers of human ERG1. Conserved amino acids are shown in solid color, ''LTGG'' residues typical of animals and fungi are shown in lighter shading, and typical plant residues are shown in outlined boxes.

Phylogenetic Analyses. One shortcoming of the above discussion is its reliance on analysis of only a few loci; therefore, we examined the total phylogeny of these enzymes in more detail. Phylogenetic trees were constructed for the amino acid sequences of SQMO and OSC (Fig. 5). All known and putative SQMO sequences to date were used; the OSC sequences then were selected for identical or comparable species.

Several main features are apparent in the trees. First, the branching pattern for each enzyme yields the same topology as a 16S rRNA gene tree. Second, each tree is consistent with a single common ancestor for each enzyme; the sequences are alignable and the pathway seems to have evolved only once. Third, the pattern of radiation of both genes is the same, i.e., the pathway is phylogenetically coherent. Any lateral transfer among species appears to have involved both genes. It may be possible to put some constraints on the history of sterol biosynthesis as a result of these observations.

Four lineages now are known to contain sterol biosynthesis genes, *Planctomycetales*, *Methylococcales*, *Myxobacteriales*, and *Eukaryota*. Vertical inheritance of sterol biosynthesis among the

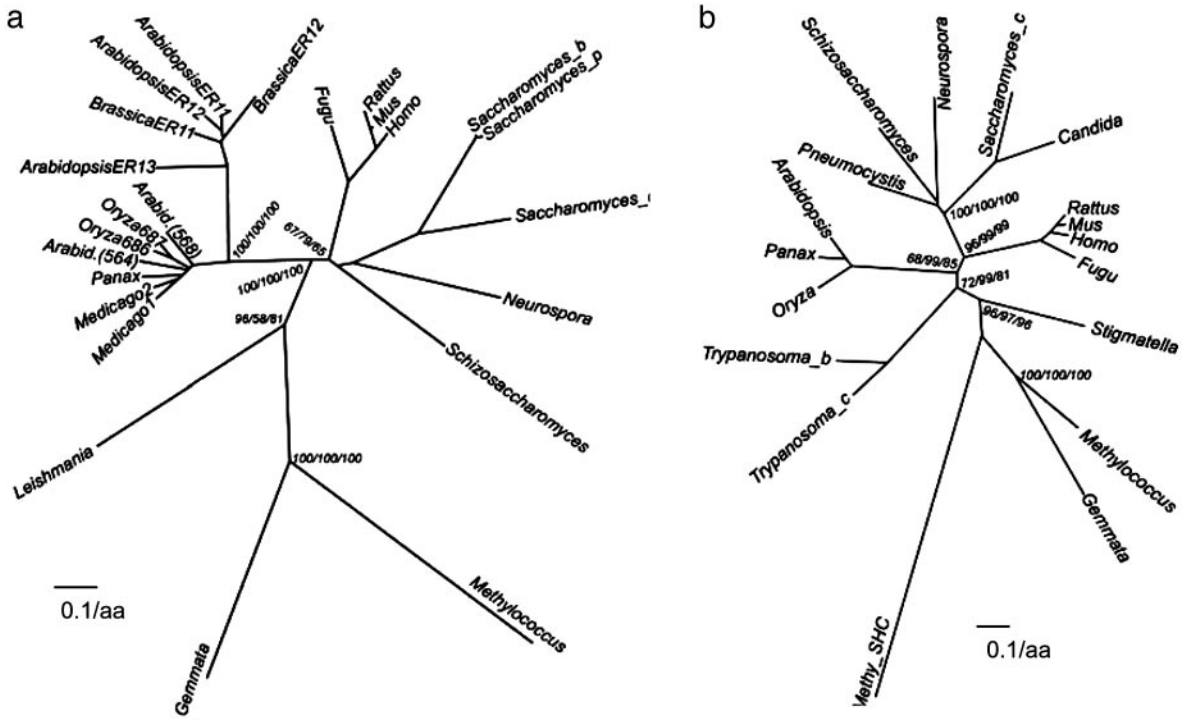

Fig. 5. (*a*) Phylogenetic trees for all known SQMO sequences of eukaryotes, plus the putative SQMO genes of *G. obscuriglobus* and *M. capsulatus*. (*b*) OSC genes corresponding to the same set of species. The OSC sequence of *S. aurantiaca* (5) also was included, despite the lack of a sequenced SQMO. Scale bars represent 0.1 changes per site. SHC of *M. capsulatus* was used as an outgroup in *b* to root the divergence of OSC from its apparent predecessor, SHC. *L. major*, *Trypanosoma brucei*, and *Trypanosoma cruzi* are unicellular protists of the group *Euglenozoa*.

three bacterial groups does not seem plausible: each is only distantly related, and a common ancestry would require selective losses from all of the other major bacterial lineages. This seems unlikely. Origin within the last common ancestor suffers from a similar problem: it would again be difficult to explain the absence of sterol synthesis genes in the majority of bacterial species. A more viable explanation for the distribution of sterol biosynthesis would be selective gene transfer between a small number of bacterial groups and an ancestral, unicellular eukaryote; it is not possible to say in which of these groups the pathway originated.

However, a recent lateral gene transfer from higher-order eukaryotes to bacteria is unlikely. Sterol biosynthesis genes of bacteria are not related closely to any extant group of eukaryotes, and the SQMO and OSC trees are parsimonious with respect to 16S rRNA phylogeny. Although the tree topology could be influenced by long-branch attraction, nothing is known about the rates of evolution of these enzymes. More persuasively, the presence of diverse sterols within the ancient geological record (2, 25) supports the divergence of sterol biosynthesis early in earth history.

Therefore, it appears most likely that bacteria and ancient eukaryotes exchanged the sterol biosynthetic pathway through gene transfer. Currently available sequence information suggests that the genes are present as a contiguous reading frame only in bacteria (Fig. 1); this would facilitate transfer of the intact pathway. Although little is known about unicellular eukaryotes at the present time, in all extant eukaryotes for which data are available, the genes are located on different chromosomes. Most examples of gene transfer between domains occur within prokaryotes [bacteria ↔ archaea; lateral transfer (12, 26)] or in the direction of prokaryotes to eukaryotes [prokaryotes → eukaryotes; lateral transfer or endosymbiotic events (27)], but transfer in any direction may have occurred among the domains before the separation of the two genes onto distinct chromosomes.

A major question remains: why do not more groups of bacteria contain sterol synthesis genes? Lateral transfer among prokaryotes is very common. Is sterol synthesis in bacteria simply a phylogenetic accident, or conversely, is it possible that the usual functional equivalent, the bacteriohopanols, may be inadequate for some species? It may be useful to assess what unusual features the three groups of sterol-synthesizing prokaryotes have. Intracellular membranes, unusual cell walls, and complex reproductive strategies are some features that are found in these organisms but are atypical of bacteria in general. Additionally, *G. obscuriglobus* is the only bacterium to possess a prototype "nucleus" (14, 18); there may be a relationship between eukaryote-like morphology and sterol synthesis. Further studies of the genome of *G. obscuriglobus* (e.g., ref. 28) may yield more information about the evolutionary significance of this unusual organism, and biochemical investigations could help illuminate the functional role(s) of sterols in both prokaryotes and eukaryotes.

Conclusions

This work illuminates the probable origin of a limited sterol biosynthesis pathway very early in earth history. The SQMO+OSC genes exist as a contiguous sequence only in bacteria, a feature that potentially facilitated lateral transfer of the pathway between bacteria and early eukaryotes. *Gemmata* is the only known organism that performs no downstream modifications of its initial C_{30} products, lanosterol and parkeol. The *Methylococcales* contain 14α- and 4α-demethylase, after which the pathway also stops. Myxobacteria produce a wider range of products, and the sterol biosynthetic capabilities of this group may be understood better if whole genome sequencing of *N. excedens* or *S. aurantiaca* is completed.

To date, it still appears that only eukaryotes produce sterols with side chain modifications at position C-24. Biomarkers extracted from the 1.64-billion-year-old Barney Creek formation include C-24 methylated and ethylated steranes (25), indicative of eukaryotes possessing advanced sterol biosynthesis. Rocks from the 2.7-billion-year-old Fortescue Group also contain these compounds (2) and, if syngenetic, push the evolution of eukaryotes back to the early Archaean. Sterol biosynthesis in bacteria may be much older still; the only apparent requirement is that it postdate the evolution of oxygenic photosynthesis.

J. Staley and C. Jenkins generously donated the Wa1-1 *Gemmata* strain and provided instructions for culturing. We thank R. Summons for helpful discussions and for the recommendation to use cycloheximide. A. Knoll, C. Cavanaugh, J. Hayes and C. Marshall read early versions of the manuscript, and we thank J. Volkman, E. DeLong, and L. Jahnke for their thorough and thoughtful reviews. D. Newman provided instruction in BLAST searching, and Z. McKinness helped with treeing methods. A. Ingalls assisted with NMR analyses, J. Ross and A. Dekas helped with the initial lipid extractions, and J. Sachs shared his gas chromatograph mass spectrometer. This work was supported by Harvard University internal funds (to A.P.) and a Harvard Junior Fellowship (to J.J.B.).

1. Volkman, J. K. (2003) *Appl. Microbiol. Biotechnol.* **60,** 495–506.
2. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. (1999) *Science* **285,** 1033–1036.
3. Bird, C. W., Lynch, J. M., Pirt, F. J., Reid, W. W., Brooks, C. J. W. & Middleditch, B. S. (1971) *Nature* **230,** 473–474.
4. Schouten, S., Bowman, J. P., Rijpstra, W. I. C. & Sinninghe Damsté, J. S. (2000) *FEMS Microbiol. Lett.* **186,** 193–195.
5. Bode, H. B., Zeggel, B., Silakowski, B., Wenzel, C. C., Reichenbach, H. & Muller, R. (2003) *Mol. Microbiol.* **47,** 471–481.
6. Kohl, W., Gloe, A. & Reichenbach, H. (1983) *J. Gen. Microbiol.* **129,** 1629–1635.
7. Tchen, T. T. & Bloch, K. (1957) *J. Biol. Chem.* **226,** 931–939.
8. Rohmer, M., Bouvier, P. & Ourisson, G. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 847–851.
9. Ourisson, G., Rohmer, M. & Poralla, K. (1987) *Annu. Rev. Microbiol.* **41,** 301–333.
10. DeSouza, N. J. & Nes, W. R. (1968) *Science* **162,** 363–364.
11. Lamb, D. C., Kelly, D. E., Manning, N. J. & Kelly, S. L. (1988) *FEBS Lett.* **437,** 142–144.
12. Boucher, Y. & Doolittle, W. F. (2000) *Mol. Microbiol.* **37,** 703–716.
13. Rohmer, M., Knani, M., Simonin, P., Sutter, B. & Sahm, H. (1993) *Biochem. J.* **295,** 517–524.
14. Lindsay, M. R., Webb, R. I., Strous, M., Jetten, M. S. M., Butler, M. K., Forde, R. J. & Fuerst, J. A. (2001) *Arch. Microbiol.* **175,** 413–429.
15. Kerger, B. D., Mancuso, C. A., Nichols, P. D., White, D. C., Langworthy, T., Sittig, M., Schlesner, H. & Hirsch, P. (1988) *Arch. Microbiol.* **149,** 255–260.
16. Strous, M., Fuerst, J. A., Kramer, E. H. M., Logemann, S., Muyzer, G., van de pas-Schoonen, K. T., Webb, R., Kuenen, J. G. & Jetten, M. S. M. (1999) *Nature* **400,** 446–449.
17. Sinninghe Damsté, J. S., Strous, M., Rijpstra, W. I. C., Hopmans, E. C., Geenevasen, J. A. J., van Duin, A. C. T., van Niftrik, L. A. & Jetten, M. S. M. (2002) *Nature* **419,** 708–712.
18. Fuerst, J. A. & Webb, R. I. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 8184–8188.
19. Brochier, C. & Philippe, H. (2002) *Nature* **417,** 244.
20. Patt, T. E. & Hanson, R. S. (1978) *J. Bacteriol.* **134,** 636–644.
21. Jahnke, L. L. (1992) *FEMS Microbiol. Lett.* **93,** 209–212.
22. Wendt, K. U., Schulz, G. E., Corey, E. J. & Liu, D. R. (2000) *Angew. Chem. Int. Ed. Engl.* **39,** 2812–2833.
23. Hart, E. A., Hua, L., Darr, L. B., Wilson, W. K., Pang, J. & Matsuda, S. P. T. (1999) *J. Am. Chem. Soc.* **121**, 9887–9888.
24. Lee, H. K., Denner-Ancona, P., Sakakibara, J., Ono, T. & Prestwich, G. D. (2000) *Arch. Biochem. Biophys.* **381,** 43–52.
25. Summons, R. E., Powell, T. G. & Boreham, C. J. (1988) *Geochim. Cosmochim. Acta* **52,** 1747–1763.
26. Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, L. D., Nelson, W. C., Ketchum, K. A., *et al.* (1999) *Nature* **399,** 323–329.
27. Doolittle, W. F. (1998) *Trends Genet.* **14,** 307–311.
28. Jenkins, C., Kedar, V. & Fuerst, J. A. (2002) *Genome Biol.* **3**, 0031.1–0031.11.