Archives of Biochemistry and Biophysics
Vol. 369, No. 1, September 1, pp. 1–10, 1999
Article ID abbi.1999.1352, available online at http://www.idealibrary.com on IDEAL®

MINIREVIEW

Cytochrome P450 and the Individuality of Species

David R. Nelson[1]
Department of Biochemistry, University of Tennessee, Memphis, Tennessee 38163

Received May 26, 1999

The P450 superfamily is expanding rapidly on many fronts. Arabidopsis genomic sequencing is producing about 2 to 3 novel P450s per week, with some clusters containing 9–14 genes. Bacterial genomes also carry surprises, such as the 20 P450s found in *Mycobacterium tuberculosis* and the 7 in *Bacillus subtilis.* The race to finish the human genome has already identified the majority of human P450s, some by expressed sequence tags only. The rapid discovery of new genes is being complemented by detailed analysis of our human genes to identify and characterize the complete set of human P450 polymorphisms and disease-causing mutations, one aspect of our "chemical individuality." Phylogenetic trees are included for plant, fungal, animal, and bacterial P450s. Emphasis is given to the higher order nomenclature of P450 clans, as a tool to see the larger picture of P450 evolution. Arabidopsis is the current record holder in P450 genes, with 186 named genes and a prediction of 350 in the total genome to be completed next year. The biosynthesis of cholesterol in bacteria is discussed in relation to CYP51 as a lanosterol 14 α-demethylase. This enzyme may have been the first eukaryotic P450. © 1999 Academic Press

Key Words: **cytochrome P450; evolution; clans; Arabidopsis; nomenclature; polymorphisms.**

BACK TO THE FUTURE

The beginning of the 20th century marked the rediscovery of Mendel's laws and set the stage for an era of genetics. This era is continuing into the 21st century with the sequencing of the human genome. The first application of Mendel's laws to human disease was made by Dr. Archibald Garrod, whose biography provided the inspiration for this essay (1). In 1902 Dr. Garrod wrote "The Incidence of Alkaptonuria: A Study in Chemical Individuality" in which he attributed alkaptonuria to a recessive genetic trait (2). Beyond his identification of the first inborn error of metabolism, Dr. Garrod is especially relevant today for his conviction " . . . that no two individuals are exactly alike chemically . . . " (1, p. 61). He considered that each person had a chemical individuality that would make that person more or less susceptible to disease. One striking passage from Bearn's biography of Garrod states that "He speculated that individual variations in the liver enzymes might influence the response of patients to a particular drug" (1, p. 142). What Garrod did not know was that his arguments for chemical individuality would resurface 80–90 years later under the guise of polymorphic variation.

Today we know that each human being is different from most other human beings at about one nucleotide in a thousand, or about 3 million sites in the genome. Only 3% of the genome is coding sequence and the frequency of polymorphisms is about 1 in 1159 bases in the coding regions of genes (3), so about 78,000 polymorphisms will occur in the coding regions. Many will be silent, but theoretically 68% of all possible mutations will result in an amino acid substitution. Therefore, we all harbor about 50,000 amino acid differences in our proteome compared with our neighbor's. This is Garrod's chemical individuality in the flesh. With 70,000 to 80,000 genes in humans, about two-thirds of our proteins will have an amino acid difference between unrelated individuals. Since there are about 50 different cytochrome P450 genes in humans (Table I), we can expect about 30 of these will bear polymorphic sites that affect the protein sequence. Some of these are already known and have been associated with an altered ability to metabolize drugs, as foreseen by Dr. Garrod (Table II). Other polymorphic sites that lie

[1] Fax: (901) 448 7360. E-mail: dnelson@utmem1.utmem.edu.

0003-9861/99 $30.00

TABLE I

Cytochrome P450s of Human, Mouse, and Rat and Their UNIGENE Entries

Human		Mouse		Rat	
CYP	UNIGENE	Cyp	UNIGENE	CYP	UNIGENE
1A1	Hs.72912	1a1	Mm.14089	1A1	Rn.10352
1A2	Hs.1361	1a2	Mm.15537	1A2	Rn.5563
1B1	Hs.154654	1b1	Mm.4443	1B1	Rn.10125
2A6, 7, 13	Hs.73864	2a4, 5	Mm.14781	2A2	Rn.9867
2A6	Hs.169233	2a12	Mm.20770	2A3	Rn.2063
2B6	Hs.1360	2b9	Mm.876	2B1, 2	Rn.2287
2B7P	Hs.110194	2b10	Mm.14177	2B3	Rn.4845
		2b13	Mm.14413		
		2b19	Mm.14098		
		2b20	No entry		
2C8	Hs.703	2c29, 39	Mm.20764	2C6	Rn.5830
2C8	Hs.166165	2c37	Mm.28533	2C7	Rn.1247
2C8	Hs.174220	2c38	No entry	2C11	Rn.10870
2C9	Hs.167529	2c40	Mm.29973	2C12	Rn.2586
2C8 17X 19	Hs.169242	2c44	Mm.26457	2C13	Rn.32070
2C18	Hs.702			2C22	Rn.10389
				2C23	Rn.2184
2D6	Hs.169876	2d9, 10	Mm.3164	2D2	Rn.32286
2D6 variant	Hs.166075	2d10, 11, 22	Mm.27803	2D3	Rn.32106
		2d12	No entry	2D4, 18	Rn.26060
		2d13	No entry	2D5	Rn.10842
		2d26	Mm.29064		
2E1	Hs.75183	2e1	Mm.13020	2E1	Rn.1372
2F1	Hs.72913	2f2	Mm.4515	2F4	Rn.10817
2G1	No entry	2g1	No entry	2G1	Rn.10909
2J2	Hs.152096	2j5	Mm.12838	2J3	Rn.10697
		2j6	Mm.6477		
		2j7	Seq confid		
		2j8	Seq confid		
		2j9	Seq confid		
2R1	Hs.16846	2r	No entry	2R1	No entry
2S1	Hs.98370	2s1, 3′ UTR	Mm.23710	2S1	No entry
3A4	Hs.45	3a11	Mm.21193	3A1, 2, 23	Rn.11291
3A5	Hs.104117	3a13	Mm.4094	3A9	Rn.10489
3A5P2	Hs.166079	3a16	Mm.30303	3A18	Rn.32085
3A7	Hs.172323				
4A11	Hs.1645	4a10	Mm.10742	4A1	Rn.5721
		4a12	No entry	4A2, 3	Rn.33492
		4a14	Mm.7459	4A8	Rn.10034
4B1	Hs.687	4b1	Mm.1840	4B1	Rn.6143
4F2	Hs.101	4f13	Mm.22045	4F1	Rn.5722
4F3	Hs.106242	4f14	Mm.10976	4F4	Rn.10170
4F8	Hs.181627	4f15	Mm.26539	4F5	Rn.10171
4F11	No entry	4f16	Mm.30504	4F6	Rn.11269
4F12	Hs.110130	4f17	No entry	4F19	Rn.21567
4F12	Hs.180570	4f18	No entry		
4X1	Hs.26040	4x	No entry	4X1	No entry
4Z1	Hs.176588	4z	No entry	4Z1	No entry
5A1	Hs.2001	5a1	Mm.4054	5A1	Rn.16283
7A1	Hs.1644	7a1	No entry	7A1	Rn.10737
7B1	Hs.144877	7b1	Mm.4781	7B1	No entry
8A1	Hs.61333	8a1	Mm.2339	8A1	Rn.10498
8B1	Hs.35718	8b1	Mm.20889	8B1	No entry
11A1	Hs.76205	11a1	Mm.28748	11A1	Rn.1401
11B1	Hs.2610	11b1	No entry	11B1, 3	Rn.32084
11B2	Hs.36986	11b2	No entry	11B2	Rn.9999
17	Hs.1363	17	Mm.1262	17	Rn.10172
19	Hs.79946	19	Mm.5199	19	Rn.21402
21	Hs.173200	21	Mm.18845	21	Rn.10225
24	Hs.89663	24	Mm.6575	24	Rn.21390

TABLE 1—*Continued*

Human		Mouse		Rat	
CYP	UNIGENE	Cyp	UNIGENE	CYP	UNIGENE
26	Hs.150595	26	No entry	26	No entry
27A1	Hs.82568	27a1	Mm.26793	27A1	Rn.34396
27B1	Hs.71210	27b1	Mm.6216	27B1	Rn.10847
39A1	Hs.20766	39	No entry	39	No entry
46	Hs.25121	46	No entry	46	No entry
51	Hs.157534	51	Mm.24155	51	Rn.6150

Note. Human UNIGENE build 76; mouse UNIGENE build 50; rat UNIGENE build 46 (April 1999).

outside the coding region may also influence gene expression. For example, CYP3A4 has a polymorphism in a nifedipine-specific response element that reduces expression compared to the wild type (4). The need for naming these polymorphisms in a consistent and curated manner has prompted the formation of a human cytochrome P450 allele nomenclature committee complete with web site and strict criteria for naming (see http://www.imm.ki.se/CYPalleles/). The information in Table II is tentative and will be greatly expanded at the CYPalleles web site.

In the discussion of natural variations in a population, it is important to distinguish between rare disease-causing mutations and polymorphisms that are present in greater than 1% of a population. Many mutations are known in P450s that cause disease, as in CYP1B1, CYP17, CYP19, CYP21, and CYP27B1, but these are different from polymorphisms that may affect drug metabolism and influence susceptibility to disease, without causing a disease directly. The difference is in the mutation frequency. Disease-causing mutations are rare, while polymorphisms by definition are not rare, but occur at 1% frequency or higher in a population. There may be a gray region just below the 1% threshold, but disease-causing mutations that do not confer a selective advantage in the heterozygous state should be considerably less frequent than 1%.

Polymorphisms can be very heterozygous, having numerous alleles, as in dinucleotide repeats, or they may be single nucleotide polymorphisms (SNPs) with only two different states and a heterozygosity near 50%. These would be easy to detect. Automated searches of the human EST database for common sequence variations have already been done and more than 3000 candidate SNPs have been discovered (24). Some rarer polymorphisms near the 1% frequency level may be difficult to find unless hundreds of chromosomes are examined. These small differences are important. Mutagenesis experiments on CYP2C2 have shown that a single amino acid substitution, S473V, allows CYP2C2 to accept progesterone as a substrate, when CYP2C2 is normally a lauric acid hydroxylase (25). The change in substrate specificity from a fatty acid to a steroid nucleus is significant. Table II is just the beginning. This characterization needs to be done for all the P450 genes in humans. Then a DNA chip or

TABLE II

A Sample of Human Cytochrome P450 Polymorphisms[a]

Allele	Mutation	Reference(s)
CYP1A1*1B	*Msp*I polymorphism associated with inducibility	5
CYP1A1*2A	I462V polymorphism	5
CYP1A2*2	F21L	6
CYP2A6*2	L160H inactive	7, 8
CYP2A6*3	Multiple changes due to gene conversion with 2A7	9
CYP2C9*2	R144C	10, 11
CYP2C9*3	I359L tolbutamide poor metabolizer	12–14
CYP2C19*2A	Splice site defect	15
CYP2C19*5A	R433W poor metabolizer of mephenytoin	16, 17
CYP2D6	65 different alleles known	18–21
CYP2E1*4	V179I no effect	22
CYP3A4*1B	Nifedipine-specific response element lowers gene expression	4
CYP3A5*2	T398N destabilizes message	23

[a] These allele names are tentative. For current details and additional alleles, see http://www.imm.ki.se/CYPalleles/.

microarray could be devised to assay the P450 profile of an individual in a single hybridization. This knowledge would characterize a person's risk of adverse drug reactions and possible predisposition to disease. This scenario is reminiscent of the opening scene in the not too futuristic movie GATTACA, where a drop of a newborn's blood is instantly analyzed and his medical future is foretold with chilling precision (26). The technology is already here to do this. In a few years Table II will be completed and the relevant P450 polymorphisms will be known. Soon policy must be set. Is P450 profiling ethical? Should the tests be done and who should have access to the information?

Cytochrome P450 has much to do with our chemical individuality. However, cytochrome P450 is equally important to our individuality as a species. No two species will have the same complement of P450 genes. Even chimpanzees and humans will not be identical in this regard. If chimpanzees have the same number of P450 genes, the sequences of these genes will have diverged over 5 million years such that the substrate specificity of the variable drug-metabolizing enzymes will be different. As mentioned above, a single base change can have dramatic effects on specificity. It is instructive to examine Table I across the three species to see the differences. Of course, the genomes of these mammals are not complete yet, so some missing genes may be found, but it is probable that the numbers of genes in subfamilies like 2C will not be the same in humans and rodents and probably not even between mice and rats.

The CYP families in Table I fall into two distinct groups: those that show no variation in family or subfamily size (CYP1, CYP5, and higher) and those that do show variation (CYP2, 3, and 4). Perhaps it is no surprise that most drug-metabolizing P450s fall into the variable group. The other P450s seem to have specific roles to play in pathways of cholesterol biosynthesis, vitamin D metabolism, bile acid metabolism, steroid biosynthesis, or thromboxane A2 biosynthesis. Some newer sequences such as CYP39 do not have known substrates, though it may be argued that they will be specific for a single substrate. This may also apply to the 1A1, 1A2, and 1B1 sequences. Since they are not in a subfamily with variable numbers of members, they may have endogenous substrates that have not been discovered yet. This seems especially likely for 1B1 since it is the cause of primary congenital glaucoma when mutated (27). CYP4B1, CYP4X1, and CYP4Z1 may also have unique endogenous substrates.

CYTOCHROME P450 AND THE ONE TRUE TREE

The era of genomics will soon answer many questions about the evolutionary history of the animal phyla and the more distant branchings in the history of life. Richard Dawkins emphasizes that there is just one true tree of life (28). We share one ancestor with mice, another with flies, and yet another single ancestor with yeast. This tree is not a statistically averaged tree, or the most parsimonious tree, but a true, one of a kind path that life has taken. As the erroneous branchings on the tree are detected with new analyses of sequence data and the morphology-based mistakes of the past are corrected, the path of P450 evolution will also emerge. We should be able to identify the one or two P450s that were ancestral to all eukaryotic P450s and move from them to the present day. We should be able to identify when new P450-dependent pathways emerged and when a CYP8B1 first appeared or a CYP11A2.

The reconstruction of P450 evolution will require in-depth sampling of species that could illuminate the paths taken. *Saccharomyces cerevisiae* (three P450s) and *Schizosaccharomyces pombe* (two P450s) are two ascomycete fungi that diverged 1000 million years ago (http://www2.bio.uva.nl/pombe/#genome). Their common ancestor had at least two P450s, CYP51 and CYP61, and probably no more than two, otherwise both lineages would have had to lose these extra genes. Candida species which are much closer to *S. cerevisiae* than *S. pombe* have a moderate CYP52 family for alkane hydroxylases (29–31), while these other fungi have not. This probably represents an expansion in Candida and not a streamlining or loss of P450s in the other species. A similar expansion occurs in Aspergillus, where at least four CYP families (CYP59, 60, 62, and 64) are used in aflatoxin biosynthesis (32–36). One must keep in mind that all living plant, animal, and fungal species today have been evolving for a billion years or more since their last common ancestor. Therefore, it may be expected that some lineages will have greatly expanded their P450 complement while others have not.

If we assume the ancestor of ascomycete fungi had CYP51 (lanosterol 14 α-demethylase) and CYP61 (22-sterol desaturase) and no other P450s, then the argument can be made that CYP61 evolved from CYP51, because CYP61 acts later in the ergosterol biosynthetic pathway. The question of when CYP61 arose can be addressed by asking when in the eukaryotic tree does ergosterol biosynthesis appear. Is it limited to fungi, or can it be seen in plants or animals, catalyzed by a CYP61 homolog? Ergosterol is widely found in fungi (37, 38). No homolog of CYP61 is known from either plants or animals, including the complete genome of *Caenorhabditis elegans.* The strongest similarity in plants is with the CYP710 family, but this is only 30%. It seems likely that CYP61 evolved only in the lineage leading to fungi, after fungi and animals diverged. This implies that the common ancestor to fungi and animals had only the CYP51 P450. Earlier ancestors including that of plants are inferred to have had just the CYP51 P450.

The argument presented above has at its core the idea that all eukaryotic P450s descended from CYP51. Furthermore, the plant, animal, and fungal lineages all evolved separate P450 collections *de novo,* starting from a CYP51. No fungal P450 family outside of CYP51 should show close resemblance to any plant or animal P450 family, and no plant family should resemble any animal family except for CYP51. This seems to be true. It has great simplifying consequences for nomenclature since these main groups can be treated independently when making trees and naming genes. Of course the lateral transfer of P450s between kingdoms must be accounted for as in the CYP55 family (39).

P450 CLANS

The nomenclature of P450s was devised to reflect the evolutionary relationships of these sequences (40). Without the simplifying assumption mentioned above, trees have been made that include P450s from plant, animal, and fungal origins. For a recent example with 107 P450 sequences see http://drnelson.utmem.edu/P450trees.html. These trees have reproducible groupings of P450s to a point, and then the more distant associations tend to be volatile. The shifting of an alignment or the addition of a new sequence can shift the deeper branches from place to place. The fungal sequences do not all cluster together nor do the plant or animal sequences. These branches seem to be interleaved. However, there is consistent assortment of these groups into a small number of kingdom-specific clusters. I believe these clusters reflect a good approximation to the true tree, as far back as we can reliably see it. Because these clusters include multiple families, a new nomenclature level has been created to name them. They are called clans. The P450 clans were introduced earlier (41) and they are refined and expanded upon here. The resemblance to clade is intentional. These clusters of P450 families are not organisms that share a single common ancestor, as in the classical definition of a clade, but they probably represent genes that diverged from a single common ancestor.

The plants have four main clans (Fig. 1). The largest has been called the group A plant P450s (42). We may now call it the plant group A clan or the 71 clan for the lowest family number in the cluster (CYP71A1 was the first cloned plant P450). Twenty-five of the 47 plant P450 families are in this clan. The other three plant clans are the 72 clan with four families CYP72, 709, 714, and 715; the 85 clan with nine families CYP85, 87, 88, 90, 702, 707, 708, 716, and 718; and the 86 clan with four families CYP86, 94, 96, and 704. CYP51, 74, 97, 710, and 711 fall outside these groups and it is not clear where they belong. The CYP74 family is anomalous because the I-helix is not conserved. This is normally one of the best conserved areas in P450s, so the 74 family always falls at the bottom of trees. It is not included here. With 358 named plant P450 sequences, the whole history of plant P450 evolution may depend on the true relationship of branching among the four plant clans and the five single families. CYP97A1 seems to cluster with the 72 clan, leaving eight groups to assemble. Comparing whole sequences will not work at this deep level of branching. It may be possible to look for other features such as intron–exon boundaries, insertions or deletions within or between structural elements, or specific motifs to further group these clusters.

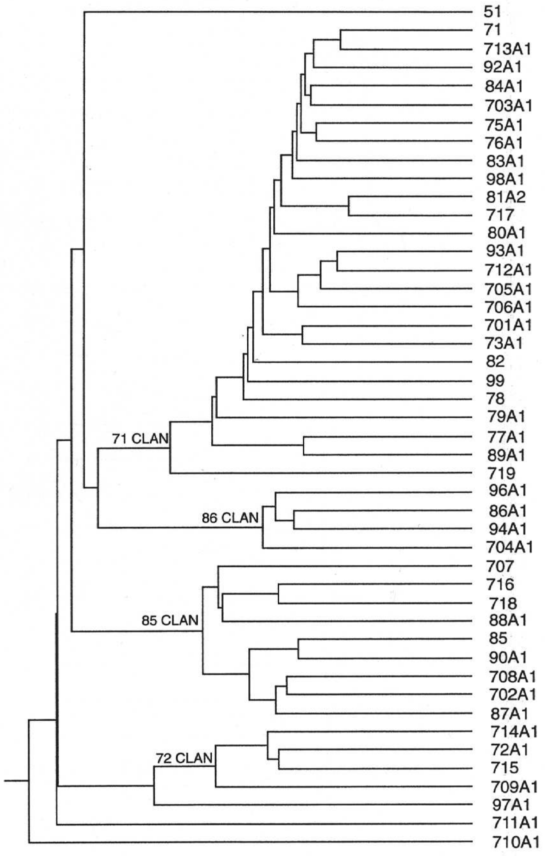

FIG. 1. A UPGMA tree of 46 plant P450s with one sequence from each family except CYP74. The tree was made using the Neighbor program in the PHYLIP package. Information on the individual sequences is available at http://drnelson.utmem.edu/CytochromeP450.html.

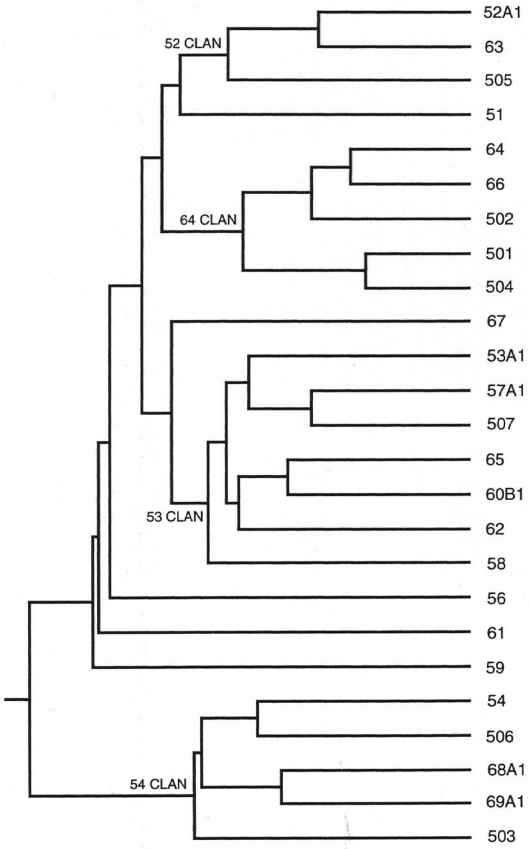

FIG. 2. A UPGMA tree of 25 fungal P450s with one sequence from each family. The tree was made using the Neighbor program in the PHYLIP package. Information on the individual sequences is available at http://drnelson.utmem.edu/CytochromeP450.html.

The fungi have four clans and a few nonaffiliated sequences (Fig. 2). The 52 clan has three families, the 53 clan has seven, the 54 clan has five, and the 64 clan has five. These sequences do not seem to have related functions. Some of them are involved in ergosterol biosynthesis (CYP51 and CYP61), others detoxify plant phytoalexins like pisatin demethylase CYP57, while others metabolize lipid carbon sources like CYP52, and still others are used to make mycotoxins as in CYP58, 59, 60, 62, 64, and 65 (32–36, 43, 44). Again, there are a small number of families that do not belong to any clans (CYP51, 56, 59, 61, and 67). P450s in fungi have been reviewed recently (45). The sampling of fungal genomes is still limited, and patterns may emerge with more sequence data.

Animals currently are assigned 43 families (Fig. 3). They cluster into several well-behaved groups with a few floaters. The clans are named for a representative family as in the 2, 3, 4, and 7 clans, a unique location as in the mitochondrial clan, or for a specific organism as in the *C. elegans* clan. The *C. elegans* clan is pure. It contains no sequences from any other organism. It seems to have derived from a common ancestor with the 2 clan, so these two groups might be considered a superclan. More than half of the *C. elegans* P450s are in this clan. The 3 clan has the mammalian 3 and 5 families, the insect 6 and 9 families, as well as the *C. elegans* 13 and 25 families. The 28 family from insects (46), the 30 family from clams (47), and the 45 family from lobster (48) are also included here. The vertebrate and invertebrate members of the 2, 3, and 4 and mitochondrial clans necessarily diverged when these major

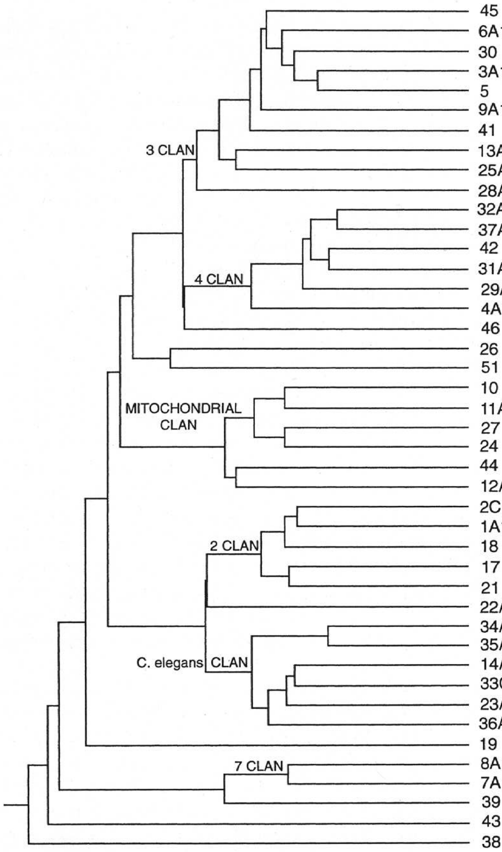

FIG. 3. A UPGMA tree of 43 animal P450s with one sequence from each family. The tree was made using the Neighbor program in the PHYLIP package. Information on the individual sequences is available at http://drnelson.utmem.edu/CytochromeP450.html.

groups split 600 million years ago. Therefore, the most distant branches within the animal clans may represent the protostome–deuterostome divergence. The recent movement of nematodes into the protostomes as part of the molting animal clade Ecdysozoa (49) means that the *C. elegans*–vertebrate divergence is included in this major dichotomy and not earlier as previously supposed. Recent phylogenetic studies suggest that there may be no intermediate phyla between the radial animals like jellyfish and the protostome–deuterostome split (50). This could leave a rather large gulf between the radial and bilateral animals. A genome project on a radial animal or a sponge would make a logical addition to the current model organism list.

The 4 clan has been greatly expanded in the insects so that it includes 24 subfamilies. Among animals, an early P450 sequence was recruited, perhaps at random, to become the genetic precursor to a radiation. All animals seem to have done this. The vertebrates expanded the 2 clan. *C. elegans* seems to have chosen this same clan. Insects selected the 4 clan and to a lesser degree the 3 clan (46, 51). There seems to be a need for 50–80 P450s in complex animals. It will be interesting to see when during evolution this occurred. Six hundred million years ago, the eukaryote ancestor of protostomes and deuterostomes must have had CYP51 and the precursors of the 2, 3, and 4 and mitochondrial clans for a minimum of five P450s. As we move further back from bilateral animals to radial animals (Cnidarians) and on to sponges, these numbers probably drop. So far, there is only one P450 known from a sponge CYP38 (GenBank Y17816). Sponges will probably have CYP51 also, but beyond that there may be very few P450s in this simple organism. As animals became more complex and needed to significantly increase their signal transduction capabilities, P450s may have been a natural choice for making and degrading signaling molecules like retinoic acid, thromboxane A2, steroids, and ecdysone. A similar process probably took place in plants as we will see with Arabidopsis.

ARABIDOPSIS, THE EVEREST OF P450-CONTAINING GENOMES

Early in the sequencing of *C. elegans,* I made a rash prediction that the worm might contain 200 P450s. That was based on a risky extrapolation and it was wrong. *C. elegans* has only 80 P450s. However, there are 186 named Arabidopsis P450 genes and 175 of these are from the genomic sequencing work. Arabidopsis is now 54.2% complete (May 3, 1999) and extrapolation is not nearly as risky. Assuming that all but 4.2% of the known sequence is from genomic sequencing, I predict there will be 175/0.5 = 350 P450 genes in Arabidopsis. There is already sequence evidence for at least 204 genes from looking at ESTs and genome survey sequences (see http://drnelson.utmem.edu/Arabfam.html for current estimates). So the old prediction that was made for *C. elegans* has already been surpassed by Arabidopsis. These Arabidopsis P450s fall into 41 families, and none of these is a clear homolog to fungal or animal P450s except the CYP51 genes (there are two CYP51s in Arabidopsis). This is in agreement with the prediction given above that all plant P450s arose starting from a CYP51 precursor.

If plants, animals, and fungi all started from ancestors with a single P450 gene at about the same time 1.2 to 1.4 billion years ago, why do some fungi have only two P450 genes, while animals have 50–80 and plants have 350? The answer is probably related to why *C. elegans* has 1049 seven transmembrane receptors while yeast has only 2 (52). There is a premium placed on specialized proteins as a species evolves complexity. If the proteins can be useful for new tasks that perhaps did not exist before, they will be duplicated and expanded to fill those needs. Multicellularity is certainly a driving force for P450 duplication in both plants and animals. The large difference between Arabidopsis and *C. elegans* or mammals probably reflects the fact that plants are more complex than animals in their biochemistry. Plants must stay put and defend their tissues from herbivorous animals (53, 54), pathogenic fungi, viruses, and bacteria. They achieve this by both physical and chemical means. The phytoalexins and toxic alkaloids plants make are the biochemical equivalent of spines and armor. The pigments of flowers are species identifiers, so they must be diverse. Both are often made with cytochrome P450 enzymes at some step or steps in their synthesis. It can be said that animals have invested heavily in senses and movement. Plants on the other hand have invested heavily in chemistry, and that is why they have more P450s than animals.

Arabidopsis is the current Everest of P450-rich genomes, but is it typical of plants? One way to evaluate this with limited data is to ask how many plant P450 families are not found in Arabidopsis. There are 358 named plant P450 genes and 186 of these are from Arabidopsis. That leaves 172 P450 genes from other plant species. Of these 172 sequences, only six plant families are not yet represented in Arabidopsis (CYP80, 92, 99, 703, 717, 719). These only include 12 P450 sequences. The CYP80 family is involved with benzylisoquinoline alkaloid biosynthesis (55, 56). The functions of the other five families are not known. From these numbers, Arabidopsis has 41/47 plant P450 families (87%) or 346/358 plant P450s have an Arabidopsis homolog (97%) and some of the missing P450 families may still be found in Arabidopsis. It follows that Arabidopsis is pretty representative of known plant P450s. One must consider that plants are

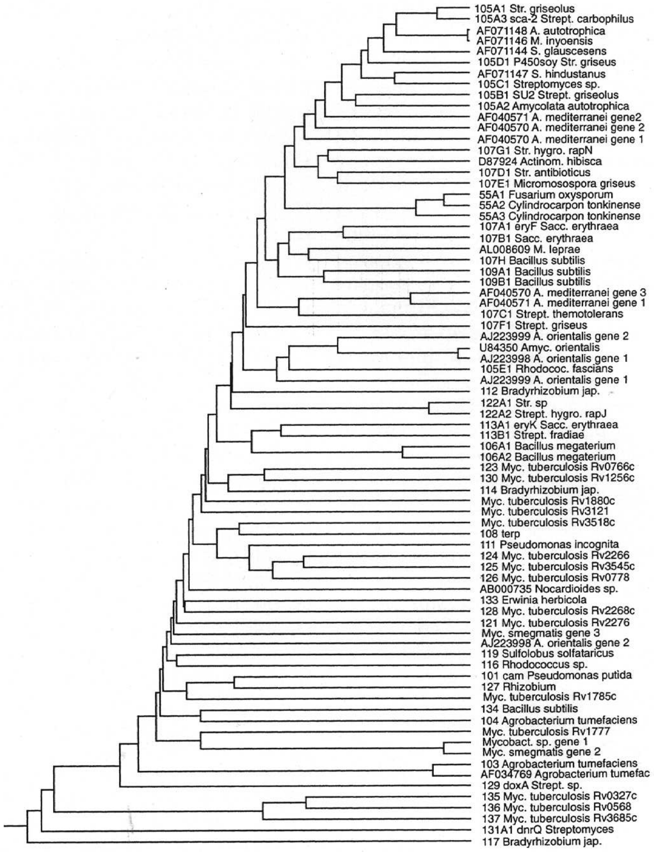
105A1 Str. griseolus
105A3 sca-2 Strept. carbophilus
AF071148 A. autotrophica
AF071146 M. inyoensis
AF071144 S. glauscesens
105D1 P450soy Str. griseus
AF071147 S. hindustanus
105C1 Streptomyces sp.
105B1 SU2 Strept. griseolus
105A2 Amycolata autotrophica
AF040571 A. mediterranei gene2
AF040570 A. mediterranei gene 2
AF040570 A. mediterranei gene 1
107G1 Str. hygro. rapN
D87924 Actinom. hibisca
107D1 Str. antibioticus
107E1 Micromosospora griseus
55A1 Fusarium oxysporum
55A2 Cylindrocarpon tonkinense
55A3 Cylindrocarpon tonkinense
107A1 eryF Sacc. erythraea
107B1 Sacc. erythraea
AL008609 M. leprae
107H Bacillus subtilis
109A1 Bacillus subtilis
109B1 Bacillus subtilis
AF040570 A. mediterranei gene 3
AF040571 A. mediterranei gene 1
107C1 Strept. themotolerans
107F1 Strept. griseus
AJ223999 A. orientalis gene 2
U84350 Amyc. orientalis
AJ223998 A. orientalis gene 1
105E1 Rhodococ. fascians
AJ223999 A. orientalis gene 1
112 Bradyrhizobium jap.
122A1 Str. sp
122A2 Strept. hygro. rapJ
113A1 eryK Sacc. erythraea
113B1 Strept. fradiae
106A1 Bacillus megaterium
106A2 Bacillus megaterium
123 Myc. tuberculosis Rv0766c
130 Myc. tuberculosis Rv1256c
114 Bradyrhizobium jap.
Myc. tuberculosis Rv1880c
Myc. tuberculosis Rv3121
Myc. tuberculosis Rv3518c
108 terp
111 Pseudomonas incognita
124 Myc. tuberculosis Rv2266
125 Myc. tuberculosis Rv3545c
126 Myc. tuberculosis Rv0778
AB000735 Nocardioides sp.
133 Erwinia herbicola
128 Myc. tuberculosis Rv2268c
121 Myc. tuberculosis Rv2276
Myc. smegmatis gene 3
AJ223998 A. orientalis gene 2
119 Sulfolobus solfataricus
116 Rhodococcus sp.
101 cam Pseudomonas putida
127 Rhizobium
Myc. tuberculosis Rv1785c
134 Bacillus subtilis
104 Agrobacterium tumefaciens
Myc. tuberculosis Rv1777
Mycobact. sp. gene 1
Myc. smegmatis gene 2
103 Agrobacterium tumefaciens
AF034769 Agrobacterium tumefac
129 doxA Strept. sp.
135 Myc. tuberculosis Rv0327c
136 Myc. tuberculosis Rv0568
137 Myc. tuberculosis Rv3685c
131A1 dnrQ Streptomyces
117 Bradyrhizobium jap.

FIG. 4. A UPGMA tree of 77 bacterial P450s with one sequence from each family, except the eukaryote-like bacterial P450s. Some sequences are not named yet. There is a problem with the nomenclature involving the CYP105 and 107 families. The tree was made using the Neighbor program in the PHYLIP package. Information on the individual sequences is available at http://drnelson.utmem.edu/CytochromeP450.html.

mostly angiosperms and the oldest angiosperm fossil is only 142 million years old (57). Monocots and dicots shared a common ancestor at about this time. Placental mammals date back about 130 million years (58), so the sequence divergence of P450s in plants should be comparable to the P450 divergence between placental and marsupial mammals. Even though plants can make a large number of specialized chemicals, there are probably very few new P450 families in plants that were not already in the monocot–dicot common ancestor. Remember, it only takes one amino acid difference to make big changes in substrate specificity (25). The test of this prediction will come with the sequence of the rice genome. Rice is a monocot and Arabidopsis is a dicot. The number of plant P450 families that are not shared between the two should give an indication of the diversity that is being missed.

BACTERIA

Bacteria are the domain of life where cytochrome P450s arose. It is not certain when or how but the evolution of P450s may predate atmospheric molecular oxygen. A phylogenetic tree of 77 bacterial P450s is shown in Fig. 4. This includes 15 of the 20 *Mycobacterium tuberculosis* P450s. The other 5 are eukaryote-like P450s and do not belong on the bacterial tree. One of these P450s is a CYP51 homolog with 33% sequence identity to human CYP51 (59). Is this protein a 14α-demethylase? Recent results using *Mycobacterium* *smegmatis* show that cholesterol is made from radiolabeled mevalonic acid in this organism (60). Therefore, CYP51 activity is required and it is probably the homolog of the *M. tuberculosis* gene identified by Yoshida and colleagues (59). Since tuberculosis bacteria are pathogens, the CYP51 gene might have been acquired from the host at some point in the past, but it could also be a descendant from the bacterial precursor for the first eukaryotic P450. This acquisition will not be resolved until other CYP51 genes can be found in bacteria and their lineage traced. A similar report of cholesterol synthesis in Staphylococcal L-forms from labeled acetate has appeared (61). The ability to find the origins of cholesterol biosynthesis in the bacteria may be coming, as more bacterial genomes are sequenced.

M. tuberculosis has 20 P450s, the largest number ever found in a single bacterial genome (62). Such an abundance of cytochromes P450 constitutes a remarkable departure from the norm. The sequences do not resemble mammalian P450s, and therefore they cannot be explained by acquisition from the host. The presence of so many P450 genes in *M. tuberculosis* gives it the potential to recruit these genes to new functions such as drug oxidations. This may be responsible for some drug resistance in this organism. Comparison of P450s sequenced in both a recent clinical isolate, CSU93, and the strain H37Rv isolated in 1905 shows only four amino acid differences. Three of the four differences involve changes in charged residues. CYP132 has the mutation L135R in the region TAA(L,R)VPG. CYP141 has the mutation K157E in the region VEP(K,E)TVH. CYP124 has two amino acid differences, D23G in the sequence PIA(D,G)IEL and Y75N in the sequence LTK(Y,N)DDV. These mutations have occurred since the last common ancestor which is a minimum of 90 years. Neither of these strains is drug resistant, so these mutations should not be responsible for a drug resistance phenotype. It may be revealing to sequence these 20 genes from a drug-resistant strain to look for significant mutations.

The functions of these P450s are not yet known. Most probably, one of them is a sterol 14α-demethylase. Some may be required for the unusual mycolic acid lipids of the cell. It is not known if any of these genes are essential for viability, but they would make potential targets for antituberculosis drugs based on P450 inhibitors.

CONCLUSIONS

The history of cytochrome P450 has gone through several stages. The age of discovery dates back to 1958 and the first spectroscopic detection of P450 in rat liver microsomes by Klingenberg (63). It continued in the 1960s with the identification of the first P450 function (64) and the indication that there were multiple forms. The age of purification arrived in the late 1960s and continued for many years. The early 1980s started the age of DNA sequencing, and a few years later the first crystal structure was solved. Now enters the age of genomics. Soon all the P450 genes in the model organisms will be catalogued. The evolutionary history of P450 will be revealed and the hard work of defining the function of hundreds of P450s will be ahead. I can hardly wait to see the 65 gene knockouts in mice, the 80 in *C. elegans,* and the 350 in Arabidopsis.

REFERENCES

1. Bearn, A. E. (1993) Archibald Garrod and the Individuality of Man, Oxford Univ. Press, Oxford.
2. Garrod, A. E. (1902) *Lancet* **ii,** 1616–1620.
3. Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S., *et al.* (1998) *Science* **280,** 1077–1082.
4. Rebbeck, T. R., Jaffe, J. M., Walker, A. H., Wein, A. J., and Malkowicz, S. B. (1998) *J. Natl. Cancer Inst.* **90,** 1225–1229.
5. Hayashi, S., Watanabe, J., Nakachi, K., and Kawajiri, K. (1991) *J. Biochem. (Tokyo)* **110,** 407–411.
6. Huang, J. D., Guo, W. C., Lai, M. D., Guo, Y. L., and Lambert, G. H. (1999) *Drug Metab. Dispos.* **27,** 98–101.
7. Yamano, S., Tatsuno, J., and Gonzalez, F. J. (1990) *Biochemistry* **29,** 1322–1329.
8. Hadidi, H., Zahlsen, K., Idle, J. R., and Cholerton, S. (1997) *Food Chem. Toxicol.* **35,** 903–907.
9. Fernandez-Salguero, P., Hoffman, S. M., Cholerton, S., Mohrenweiser, H., Raunio, H., Rautio, A., Pelkonen, O., Huang, J. D., Evans, W. E., Idle, J. R., *et al.* (1995) *Am. J. Hum. Genet.* **57,** 651–660.
10. Rettie, A. E., Wienkers, L. C., Gonzalez, F. J., Trager, W. F., and Korzekwa, K. R. (1994) *Pharmacogenetics* **4,** 39–42.
11. Crespi, C. L., and Miller, V. P. (1997) *Pharmacogenetics* **7,** 203–210.
12. Sullivan-Klose, T. H., Ghanayem, B. I., Bell, D. A., Zhang, Z.-Y., Kaminsky, L. S., Shenfield, G. M., Miners, J. O., Birkett, D. J., and Goldstein, J. A. (1996) *Pharmacogenetics* **6,** 341–349.
13. Aithal, G. P., Day, C. P., Kesteven, P. J., and Daly, A. K. (1999) *Lancet* **353,** 717–719.
14. Haining, R. L., Hunter, A. P., Veronese, M. E., Trager, W. F., and Rettie, A. E. (1996) *Arch. Biochem. Biophys.* **333,** 447–458.
15. de Morais, S. M. F., Wilkinson, G. R., Blaisdell, J., Nakamura, K., Meyer, U. A., and Goldstein, J. A. (1994) *J. Biol. Chem.* **269,** 15419–15422.
16. Xiao, Z. S., Goldstein, J. A., Xie, H. G., Blaisdell, J., Wang, W., Jiang, C. H., Yan, F. X., He, N., Huang, S. L., Xu, Z. H., and Zhou, H. H. (1997) *J. Pharmacol. Exp. Ther.* **281,** 604–609.
17. Ibeanu, G. C., Blaisdell, J., Ghanayem, B. I., Beyeler, C., Benhamou, S., Bouchardy, C., Wilkinson, G. R., Dayer, P., Daly, A. K., and Goldstein, J. A. (1998) *Pharmacogenetics* **8,** 129–135.
18. Gough, A. C., Miles, J. S., Spurr, N. K., Moss, J. E., Gaedigk, A., Eichelbaum, M., and Wolf, C. R. (1990) *Nature* **347,** 773–776.
19. Saxena, R., Shaw, G. L., Relling, M. V., Frame, J. N., Moir, D. T., Evans, W. E., Caporaso, N., and Weiffenbach, B. (1994) *Hum. Mol. Genet.* **3,** 923–926.

20. Broly, F., Marez, D., Lo Guidice, J.-M., Sabbagh, N., Legrand, M., Boone, P., and Meyer, U. A. (1995) *Hum. Genet.* **96,** 601–603.

21. Marez, D., Legrand, M., Sabbagh, N., Guidice, J. M., Spire, C., Lafitte, J. J., Meyer, U. A., and Broly, F. (1997) *Pharmacogenetics* **7,** 193–202.

22. Fairbrother, K. S., Grove, J., de Waziers, I., Steimel, D. T., Day, C. P., Crespi, C. L., and Daly, A. K. (1998) *Pharmacogenetics* **8,** 543–552.

23. Jounaidi, Y., Hyrailles, V., Gervot, L., and Maurel, P. (1996) *Biochem. Biophys. Res. Commun.* **221,** 466–470.

24. Buetow, K. H., Edmonson, M. N., and Cassidy, A. B. (1999) *Nat. Genet.* **21,** 323–325.

25. Ramarao, M., and Kemper, B. (1995) *Mol. Pharmacol.* **48,** 417–424.

26. Silver, L. (1997) *Nat. Genet.* **17,** 260–261.

27. Stoilov, I., Akarsu, A. N., Alozie, I., Child, A., Barsoum-Homsy, M., Turacli, M. E., Or, M., Lewis, R. A., Ozdemir, N., Brice, G., Aktan, S. G., Chevrette, L., Coca-Prados, M., and Sarfarazi, M. (1998) *Am. J. Hum. Genet.* **62,** 573–584.

28. Dawkins, R. (1996) *in* The Blind Watchmaker, pp. 255–284, Norton, New York.

29. Sanglard, D., and Loper, J. C. (1989) *Gene* **76,** 121–136.

30. Ohkuma, M., Muraoka, S., Tanimoto, T., Fujii, M., Ohta, A., and Takagi, M. (1995) *DNA Cell Biol.* **14,** 163–173.

31. Ohkuma, M., Zimmer, T., Iida, T., Schunck, W. H., Ohta, A., and Takagi, M. (1998) *J. Biol. Chem.* **273,** 3948–3953.

32. Keller, N. P., Kantz, N. J., and Adams, T. H. (1994) *Appl. Environ. Microbiol.* **60,** 1444–1450.

33. Yu, J., Chang, P.-k., Cary, J. W., Wright, M., Bhatnagar, D., Cleveland, T. E., Payne, G. A., and Linz, J. E. (1995) *Appl. Environ. Microbiol.* **61,** 2365–2371.

34. Yu, J., Chang, P. K., Cary, J. W., Bhatnagar, D., and Cleveland, T. E. (1997) *Appl. Environ. Microbiol.* **63,** 1349–1356.

35. Brown, D. W., Yu, J. H., Kelkar, H. S., Fernandes, M., Nesbitt, T. C., Keller, N. P., Adams, T. H., and Leonard, T. J. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1418–1422.

36. Yu, J., Chang, P. K., Ehrlich, K. C., Cary, J. W., Montalbano, B., Dyer, J. M., Bhatnagar, D., and Cleveland, T. E. (1998) *Appl. Environ. Microbiol.* **64,** 4834–4841.

37. Weete, J. D., and Gandhi, S. R. (1997) *Lipids* **32,** 1309–1316.

38. Parks, L. W., and Casey, W. M. (1995) *Annu. Rev. Microbiol.* **49,** 95–116.

39. Tomura, D., Obika, K., Fukamizu, A., and Shoun, H. (1994) *J. Biochem. (Tokyo)* **116,** 88–94.

40. Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. J., Feyereisen, R., Waxman, D. J., Waterman, M. R., Gotoh, O., Coon, M. J., Estabrook, R. W., Gunsalus, I. C., and Nebert, D. W. (1996) *Pharmacogenetics* **6,** 1–42.

41. Nelson, D. R. (1998) *Comp. Biochem. Physiol. C* **121,** 15–22.

42. Durst, F., and Nelson, D. R. (1995) *Drug Metab. Drug Interact.* **12,** 189–206.

43. Trapp, S. C., Hohn, T. M., McCormick, S., and Jarvis, B. B. (1998) *Mol. Gen. Genet.* **257,** 421–432.

44. Alexander, N., Hohn, T. M., and McCormick, S. P. (1998) *Appl. Environ. Microbiol.* **64,** 221–225.

45. Vanden Bossche, H., and Koymans, L. (1998) *Mycoses* **41**(Suppl. 1), 32–38.

46. Danielson, P. B., Foster, J. L., McMahill, M. M., Smith, M. K., and Fogleman, J. C. (1998) *Mol. Gen. Genet.* **259,** 54–59.

47. Brown, D. J., Clark, G. C., and Van Beneden, R. J. (1998) *Comp. Biochem. Physiol. C* **121,** 351–360.

48. Snyder, M. J. (1998) *Arch. Biochem. Biophys.* **358,** 271–276.

49. Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997) *Nature* **387,** 489–493.

50. Adoutte, A., Balavoine, G., Lartillot, N., and de Rosa, R. (1999) *Trends Genet.* **15,** 104–108.

51. Feyereisen, R. (1999) *Annu. Rev. Entomol.* **44,** 507–533.

52. Bargmann, C. I. (1998) *Science* **282,** 2028–2033.

53. Schuler, M. A. (1996) *Plant Physiol.* **112,** 1411–1419.

54. Gonzalez, F. J., and Nebert, D. W. (1990) *Trends Genet.* **6,** 182–186.

55. Kraus, P. F., and Kutchan, T. M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 2071–2075.

56. Pauli, H. H., and Kutchan, T. M. (1998) *Plant J.* **13,** 793–801.

57. Sun, G., Dilcher, D. L., Zheng, S., and Zhou, Z. (1998) *Science* **282,** 1692–1695.

58. Kumar, S., and Hedges, S. B. (1998) *Nature* **392,** 917–920.

59. Aoyama, Y., Horiuchi, T., Gotoh, O., Noshiro, M., and Yoshida, Y. (1998) *J. Biochem. (Tokyo)* **124,** 694–696.

60. Lamb, D. C., Kelly, D. E., Manning, N. J., and Kelly, S. L. (1998) *FEBS Lett.* **437,** 142–144.

61. Hayami, M., Okabe, A., Sasai, K., Hayashi, H., and Kanemasa, Y. (1979) *J. Bacteriol.* **140,** 859–863.

62. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., III, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Barrell, B. G., *et al.* (1998) *Nature* **393,** 537–544.

63. Klingenberg, M. (1958) *Arch. Biochem. Biophys.* **75,** 376–386.

64. Estabrook, R. W., Cooper, D. Y., and Rosenthal, O. (1963) *Biochem. Zeit.* **338,** 741–755.