Cytochrome P450 Nomenclature

David R. Nelson

1. Introduction

1.1. Approaching 1000

With genome projects spewing forth DNA sequence at tens of Megabases per year, the problem of genetic nomenclature becomes daunting. In the field of cytochrome P450 (P450), there are more than 750 sequences and they are accumulating rapidly. At the current rate, there will be more than 1000 P450 sequences by late 1998 or early 1999. For these data to be accessible and useful they must be sorted and categorized in some meaningful way. To that end, a cytochrome P450 nomenclature system was devised ***(1)***. This system relied on evolutionary relationships as depicted in phylogenetic trees or dendrograms derived from the P450 protein sequences. The whole collection of sequences represents the P450 superfamily, with families and subfamilies being arbitrarily defined as distinct clusters on the tree. Originally, 40% identity was used as a cutoff for family membership and 55% was used for subfamily membership. Since the inception of the system, both numbers have crept downward. The actual decision to include a sequence in an existing group largely depends on how it clusters on a tree and not so much on the absolute percent identity, which is more or less a rule of thumb. The most recent published compilation of the nomenclature is given in Nelson et al. (1996) ***(2)***. More up-to-date information is available at http://drnelson.utmem.edu/nelsonhomepage.html.

1.2. Looking to the Future

This system has been adopted by the P450 research community and survived intact for <10 yr. Some cumbersome features have been changed (Roman numerals were replaced by Arabic numbers) but the scheme is mainly as it was in 1987. There have been some growing pains associated with large numbers

From: *Methods in Molecular Biology, Vol. 107: Cytochrome P450 Protocols*
Edited by: I. R. Phillips and E. A. Shephard © Humana Press Inc., Totowa, NJ

of sequences. One of the worst of these is the lack of two digit numbers for new P450 families. The original scheme reserved blocks of numbers for different groups of organisms. CYP families 1–49 were reserved for animals, families 51–69 for lower eukaryotes, 71–99 for plants, and 101 and up for bacteria. The plant P450 field has exploded recently and family numbers for plants in the two digit range have already run out. Lower eukaryotes are soon going to have the same problem, with animals not far behind. The solution has been to assign blocks of three digit numbers to accommodate the overflow. Lower eukaryotes will be assigned numbers in the 501–699 range. Plants will take 701–999. Animals will have to jump to the 301–499 range because bacteria are already in the lower hundreds, 101–299. It is conceivable that three digit numbers will still not be enough, but it should hold us well into the next decade.

2. The Process of Assigning a Name

New P450 sequences frequently appear in Genbank. It is a time-consuming process to search Genbank and sort out new sequences from those that have already been named. This is especially difficult for the expressed sequence-tags (ESTs), because the protein sequence has to be derived from the nucleotide sequence and there are many frameshifts in the EST data. Highest priority goes to sequences submitted by researchers for naming. These sequences undergo an initial scan that aligns them against the appropriate subset of P450 sequences. The alignment algorithm is part of the MacVector package of sequence analysis software (IBI). A new sequence can be aligned with a folder containing about 200 sequences in less than a min. The folders contain bacterial (55), plant (200), insect (104), CYP1, 2, 17, and 21 (168), and other (236) sequences, a combined total of 763 sequences. A few quick scans identifies the best match to a known P450 sequence. The new sequence is then added to a master alignment containing the best match. The revised alignment is converted to an integer array 730 positions long by the number of sequences in the alignment. A 200 sequence alignment in integer form occupies 584 kb.

A program called COMPARE is used to compare the new sequence to every other sequence in the array. The percent identity is given for each sequence. Often, a name can be assigned based on these numbers alone. If a sequence is 85% identical to a known sequence, then it will clearly fall into the same subfamily and a name can be given without further analysis. If the best matches are less obvious, say in the 40–50% range, then it is usually necessary to construct a tree. A difference matrix is computed from the array by comparing every sequence with every other sequence, a process that can take a while for a large array. The difference matrix is stored and used to compute a tree by the UPGMA method. There is the option to use several different scoring matrices in computing the difference matrix, but, for simplicity, it is usually best to use a unit matrix that only counts identities. The numbers are later adjusted for multiple mutations at the same site as described in Nelson and Strobel *(3)*.

A tree will show the relationships between sequences better than the numbers from the COMPARE program. Examination of the tree in addition to the COMPARE output is sufficient to assign a name. Some sequences naturally fall midway between existing families or subfamilies. These sequences are difficult to assign, because they are sensitive to small changes in the alignment or to the choice of scoring matrix. If two different outcomes are found with the corrected unit matrix data or a PAM250 tree, choose the PAM250 tree over the unit-matrix tree. This has been the case with the bacterial CYP105 family. Sometimes this is just a judgment call and the result could go either way. The results are partly historical. For example, the CYP2D subfamily is quite different from the other CYP2 family members. When CYP2D1 was first reported, there was a fairly large gap between the CYP2 family and the CYP1 family. Some, but not all, of the 2D sequences were more than 40% identical to other CYP2 family members. At the time it seemed prudent to keep the 2D cluster within the CYP2 family. As more sequences have been reported, the break between the CYP1 and CYP2 families is less clear, and it is now apparent that the 2D subfamily should have been given full family status.

Because of the large number of factors that go into assigning a name and the possible involvement of confidential sequences, the Committee on Standardized Cytochrome P450 Nomenclature asks that individuals do not name their own P450 genes. This will avoid the possibility of assigning identical names for different sequences. Instead, individuals are encouraged to request help in naming a new gene, by submitting the sequence by e-mail or fax. The appropriate addresses are given on the P450 web page at http://drnelson.utmem.edu/nelsonhomepage.html. These sequences will be held in confidence until they are published or appear in Genbank. The name that is assigned will be posted as well as the species that it is from and the name of the person or lab submitting the sequence. If the sequence is already in the database, the individual who submitted the sequence will be notified. In the case of Arabidopsis, where there are about 175 EST sequences for P450s, the matching ESTs may be cross-referenced as time permits. If one has concerns about patents and does not want the sequence he or she is working on to be known, then that individual should not submit the sequence for naming. Private names, for sequences without species given or with explicit requests to withhold EST cross-references, will not be honored. A nomenclature cannot be secretive. If that is not compatible with one's wishes, then that individual must hold his or her sequence back until these matters have been resolved.

3. Benchmarks for P450 Sequence Alignment

The construction of a tree is dependent on the sequence alignment and also on what parts of a sequence are included in the analysis. This is becoming even more clear as phylogenies are compared from different sets of sequence data and even nonsequence data using cladistics. In a comparison of mitochondrial genes coding for 13 different protein sequences over a wide range of organisms, analysis of all the sequence data gave the wrong tree, with good statistical support ***(4)***. Using just hydrophobic amino-acid codons gave poor results; hydrophilic and charged amino-acid codons gave better results. The best tree, which was in agreement with expected relationships among these organisms, was produced using only the parts of the genes that coded for important three-dimensional features in these proteins. These findings have implications for P450 sequence alignments and tree building for nomenclature purposes. More reliable results may be obtained by leaving out the poorly conserved regions of the sequences that are the most difficult to align. The regions remaining should include the sequence benchmarks that are used to align the sequences.

These benchmarks are few in the N-terminal region, and more abundant in the C-terminal from the I-helix to the end of the P450 molecule. Many of these benchmark regions are specific for families or higher taxonomic groups, like bacterial or mitochondrial P450s that tend to cluster together. When aligning difficult sequences with low similarity to other sequences, pay particular attention to these benchmark regions. Alignment algorithms cannot use this information, and they are often quite far off when it comes to distantly related P450 sequences. This is understandable when a bacterial sequence may be only 8–9% identical to a mammalian sequence.

At the N-terminal of many plant and animal P450s is a proline-rich sequence. It is often PPGP, but it may be (hydrophobic; hdr)PGP. This sequence is usually followed four residues later by P(hdr)(hdr)G(polar)(hdr) as in PIIGNL. This region follows the N-terminal membrane-anchor sequence, which is usually not conserved and cannot be aligned with any certainty. About 20 amino-acid residues later, there is a common sequence, KYG or RYG. This tripeptide is present in CYP51, lanosterol-14α-demethylase, which is one of the more ancient eukaryotic P450s involved in cholesterol biosynthesis. It is also present in CYP7, CYP8, and CYP19 sequences, but not in CYP4 or CYP52 (fatty-acid and alkane hydroxylases) nor in bacterial-like sequences. The two basic P450 forms are called E-like and B-like P450s for eukaryote and bacterial-like, respectively. The ancient origin of the KYG motif is further confirmed by its presence in CYP110 from *Anabaena*, a cyanobacterium. In modified form (ELG) it is also seen in CYP11B and CYP102 ($P450_{BM-3}$) from *Bacillus megaterium*, where it occurs just before the β 1-1 strand. Both CYP110 and

CYP102 are E-like bacterial P450s, suggesting the fundamental split between E-like and B-like P450 structures occurred before eukaryotes developed. One E-like P450 crystal structure is known of CYP102, and several B-like P450 crystal structures are known of CYP101 ($P450_{cam}$) and CYP108 ($P450_{terp}$). For a comparison of these structures, see Peterson and Graham-Lorence *(5)*.

In bacterial B-like P450s, near amino acid 50 there is a conserved trp residue, often seen in the pattern WXXT(R,K), as in WIATK or WLVTR. This is the first region to look for near the N-terminus of a new bacterial sequence. About 100 amino-acid residues from the N-terminus of bacterial sequences, there is another region, around the C-helix, that is critical when aligning bacterial sequences with eukaryotic sequences. The motif is DPPXHXXXR. This motif corresponds to a C-helix sequence, WXXXR, conserved in most eukaryotic P450s. The R is often followed by another basic group as in WREQRR. Some lower eukaryotic P450s (CYP53, 57, and 58) have the H instead of the W seen in most eukaryotes.

After the C-helix, there is a long nonconserved stretch with few alignment cues. One worth mentioning is the (N,D,S)(V,I,T)(V,I) sequence around positions 175–180 in eukaryotic sequences. This E-helix region can be diagnostic for some families, as in the 3A subfamily, where the sequence reads GAYSMD(V,I)I, or the 4A subfamily SLMTLDT(I,V). After this region, the next well-conserved area is the I-helix. From here (near position 300) the P450s are much more strongly conserved and alignments are much easier. The I-helix has a characteristic sequence A(A,G)X(E,D)T, where T is part of the molecular oxygen-binding site. This region is not conserved in enzymes like allene oxide synthase (CYP74) that use peroxides as substrates, because they do not have to bind molecular oxygen. Consequently, the CYP74 family is one of the most distant branches on the P450 tree. This may not reflect its true evolutionary history, because one of the key regions used in computing trees is not conserved.

Exactly 16 amino-acid residues from the T previously mentioned is a conserved P. This P is present even in many bacterial sequences so it is a good marker for the junction between the I and J helices. The sequence conservation continues for another 15 residues in eukaryotes to a highly conserved G or P residue. At this point, there seems to be some sequence length variation in the region between the J-helix and the K-helix. The K-helix is the best conserved feature in P450s, with its invariant EXXR charge pair. About 22 amino-acid residues beyond the K-helix R, there is an (aromatic)X(I,V,L)P(K,A)G sequence that spans the connection between the β-2-2 strand and the β-1-3 strand (nomenclature of Peterson and Graham-Lorence, **ref. 5**). The P(K,A) pair lies in this gap and causes a sharp change in direction. The region beyond this area is not structurally well-defined, but 17 residues from the G there is a sequence, (aromatic)XX(P,D), that is very helpful in making alignments. It is followed 4–5 residues later by the PERF sequence, with some variations as in PSRF, PDNF, PQRW, and so on. The PERF motif is not present in bacterial B-like P450 sequences, though it is in CYP102, CYP110, and CYP118 bacterial E-like P450s. It is part of the meander before the heme thiolate ligand (nomenclature of Peterson and Graham-Lorence, **ref. *5***). After this point, there are length variations before reaching the signature sequence of all P450s, FXXGXXXCXG, with some variations allowed except at Cys. Beyond the signature, 18 residues from the Cys, there is a conserved tetrapeptide, LQNF, or variants of it. The C-terminal region is quite variable with an unexplained occurrence of a PR approx 33–34 residues from the F in LQNF. In many sequences, these are the last two amino-acid residues, whereas in others there is an extension beyond the PR sequence.

These form the main alignment benchmarks that are used in making P450 sequence alignments. Often the distance between benchmarks is absolutely the same and a cursory check for reasonable homologies between the segments can confirm that the alignment is probably correct without introducing gaps. Of course, this is most difficult to do in the N-terminal region and when aligning bacterial or lower eukaryotic sequences that are very divergent from the other sequences.

4. Whole Genome Collections: Yeast, *C. elegans*, and Humans

Subfamilies of P450 genes were thought to represent tightly linked clusters of genes ***(6)***. The nomenclature should then reflect physical proximity on chromosomes. At one point, we considered revising the nomenclature when these clusters were sequenced to accurately reflect the sequence order of genes within a given cluster. This approach was not implemented, and it seems to have been based on erroneous assumptions. The genome projects are beginning to show that sequences from the same subfamily are not always linked. In *C. elegans*, the CYP13A subfamily exists on two different chromosomes, and this is also true for the CYP25 family. Therefore, the notion of tightly clustered subfamilies is at least partially incorrect. Sequencing of genomic DNA also shows that P450s from different subfamilies can occur mixed in the same cluster. Human chromosome 19q13.2 contains the following linear arrangement of P450 genes: *CYP2F1*, *CYP2A6*, *CYP2A7P*, *CYP2B7P*, *CYP2B6*, *CYP2A12*, and *CYP2F1P*, as shown on the Lawrence Livermore National Laboratories genome project home page at http://www-bio.llnl.gov/genome/html/gene_ideogram.html. Clearly the subfamilies identified by sequence analysis are not discrete clusters in the genome. Chromosomal evolution seems to be much more rapid than gene evolution, leading to the scrambling of gene clusters on a single chromosome and between chromosomes.

The most completely sequenced animal genome is that of *C. elegans*. When the sequence of this genome was about 62% complete (April, 1997), an extensive search for P450 genes identified 70 different sequences, leading to an estimate of about 100 P450 genes in *C. elegans*. For a list of these sequences, *see* http://drnelson.utmem.edu/CEP450s.html. For translations of the genes, *see* http://drnelson.utmem.edu/CEP450spep.html. This genome project should be finished by the end of 1998, so a complete set of P450s from an animal will then be available. There is a great increase in the number of P450s from yeast to *C. elegans*. The whole yeast genome has only three P450s, CYP51, CYP56, and CYP61, whereas animals have about 30 times that number. Their function is clearly not required for a eukaryotic existence, except to make cholesterol. They seem to be more important for the multicellular state.

Human P450 genes are best represented in the dbEST section of Genbank. A search through about 800 ESTs in December 1995 identified 12–14 sequences that were as yet unknown P450s in humans. Since then, two or three of these have appeared in the database as newly identified P450 genes, but there are still about 10 that have yet to be described. The author's estimate at that time for human P450s was about 54. That assumed 83% coverage of all human genes in the EST database. Since then, the extent of coverage has been revised to 50–60%. Therefore, the estimate of human P450 genes was probably low and it may be closer to 75–90, similar to the number in the *C. elegans* genome. It should not be assumed that the same genes will be present in both organisms. So far, there do not appear to be any mitochondrial P450s or steroidogenic P450s in *C. elegans*; however, a CYP51 gene fragment is present, and the *CYP29* and *31* families are most similar to *CYP4*.

A new search of the EST database is warranted, as the number of entries has more than doubled since 1995 (301,000–709,000). The sheer numbers of hits makes this search a time consuming and difficult task. In 1995 the query (p450 AND human[LIB]) NOT reductase gave 1375 hits. One problem that soon became apparent while looking at these data was the presence of alu repeats. *CYP2D6* and *CYP1A2* have alu repeats, so every alu repeat in dbEST had the potential to give a spurious P450 hit in a blast search. By adding NOT alu to the query, the number of hits dropped from 1375 to 787. The same query done on May 20, 1997 (without [LIB], owing to changes in the database organization) gave 1635 hits. Clearly, this will take a major commitment to sort through, but the rewards may be identification of nearly all expressed human P450 genes.

5. The Problems of Short ESTs or PCR Products

The dbEST section of Genbank as of May 16, 1997 has 709,000 human ESTs, 182,000 mouse ESTs, and 31,000 Arabidopsis ESTs. A search of this database http://www2.ncbi.nlm.nih.gov/dbST/dbest_query.html on July 9, 1996 for *Arabidopsis* P450 matches by the following expression (P450 AND Arabidopsis[LIB]) NOT reductase gave 163 fragments. These have all been translated and compared to the plant database that the author had at that time. Although some of the fragments could not be identified as P450s, most were assigned to specific sequences or families. These are listed in the P450 Web page at http://drnelson.utmem.edu/ArabESTS.html. An identical search done on May 19, 1997 gave 227 ESTs, 64 more than the previous year. However, 48 of these were matches to a tobacco sequence, D64052, that shows a P450 spectrum and P450 activity, but does not appear to be a P450 family member based on descent from a common ancestor *(7)*. In addition, all of these ESTs had a high probability of being ribosomal RNAs, so these 48 sequences probably do not represent P450 sequences. Only 14 or 15 of the new ESTs that appeared since July, 1996 seem to be legitimate P450 ESTs.

When these sequences are translated and aligned with existing full-length sequences and other ESTs, they can be assigned to families only if they are very similar to existing sequences. If they are only moderately similar they cannot be named because the percent identity will change when full-length sequences become known. The N-terminal region of P450 sequences is less conserved than the C-terminus, so a polymerase chain reaction (PCR) fragment or an EST fragment that is from the C-terminal region and is about 55% identical to a known sequence may have less than 40% identity when the whole sequence is known. The 40% value is the approximate cutoff for inclusion in a P450 family. To make the problem worse, some N-terminal fragments may be from the same gene as a C-terminal fragment in the collection, but it is not possible to know this without the overlapping region. Therefore, naming of short fragments is dangerous and may lead to multiple names for the same gene.

Some laboratories have been using degenerate PCR quite successfully to amplify and sequence dozens of P450 fragments from the same species or related species, and they would like to have these named. One lab amplified the region from the I-helix to the heme signature region and sent the sequences of dozens of insect PCR fragments to be named. Another lab did the same from the heme signature region to the end of the P450 gene and submitted the sequences of more than 100 insect DNA fragments. The two groups' fragments do not overlap, making it impossible to compare them with each other. The I-helix to heme signature fragments were longer and more informative, and they could be used to assign names, though it may be necessary to revise these later as the full-length sequences are determined.

The short fragments of P450 sequences cannot be used in tree building with complete sequences, because the results are very misleading. The trees that result from such mixtures of short and long sequences will probably cluster the short fragments correctly with their closest related sequence. However, that longer sequence may now be shifted to a completely wrong location on the tree. This happens because the difference matrix does not make adjustments for short sequences. All percent identities are treated as if they were from same length comparisons. This would be acceptable if the percent identity were nearly constant over the length of P450s, but it is not. One solution to this problem is to make trees from segments of the sequence alignment. If all there is to work with is C-terminal fragments, such as the PCR fragments from the I-helix to the heme signature, then use just this region from all the sequences, including full-length sequences. Another possibility used by Rene Feyereisen, is to make an estimation of the percent identity for the full-length sequence based on a polynomial regression between percent identities for full-length sequences vs percent identities for the desired region from the same sequences *(8)*. Both methods result in difference matrices that are adjusted to same length sequences.

6. Weaknesses of the Nomenclature System

The nomenclature system depends on evolutionary relationships between similar sequences being correctly or at least consistently determined. The mechanisms for doing this have improved over the last 10 yr, so that there are better algorithms than UPGMA and better scoring matrices than a unit matrix adjusted for multiple hits at a given site. The neighbor-joining method of Saitou and Nei *(9)* is generally considered better at reconstructing phylogenies than UPGMA, and the BLOSUM 62 matrix *(10)* is the usual default in many applications today. The trees built from P450 sequences would be better if these newer methods were used. The drawback of making this change is the potential to significantly alter the existing nomenclature, which has been developed over a 10-yr period. Perhaps this should be done anyway to achieve the benefits of more accurate tree-building methods. There is a good probability that most of the branching pattern changes would be at deep nodes on the trees and these would not affect the nomenclature, because family assignment is based on identities of ~35–40% or greater. The clusters at these levels should not change much no matter what tree-building methods are used. This was seen in a comparison of P450 trees made by neighbor-joining and UPGMA programs before the nomenclature system was adopted *(3)*. Only 1 out of 34 branches changed position and this would not have altered the nomenclature. In the future, a switch should be made, if not too disruptive, to more sophisticated tree-making algorithms and scoring matrices.

References

1. Nebert, D. W., Adesnik, M., Coon, M. J., Estabrook, R. W., Gonzalez, F. J., Guengerich, F. P., Gunsalus, I. C., Johnson, E. F., Kemper, B., Levin, W., Phillips, I. R., Sato, R., and Waterman, M. R. (1987) The P450 gene superfamily: recommended nomenclature. *DNA* **6,** 1–11.
2. 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) P450 Superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* **6,** 1–41.
3. Nelson, D. R. and Strobel, H. W. (1987) Evolution of cytochrome P-450 proteins. *Mol. Biol. Evol.* **4,** 572–593.
4. Balter, M. (1997) Morphologists learn to live with molecular upstarts. *Science* **276,** 1032–1034.
5. Peterson, J. A., and Graham-Lorence, S. E. (1995) Bacterial P450s. Structural similarities and functional differences, in *Cytochrome P450: Structure, Mechanism and Biochemistry*, 2nd ed. (Ortiz de Montellano, P. R, ed.), Plenum, NY, pp. 151–180.
6. Nelson, D. R., Kamataki, T., Waxman, D. J., Guengerich, F. P., Estabrook, R. W., Feyereisen, R., Gonzalez, F. J. Coon, M. J., Gunsalus, I. C., Gotoh, O., Okuda, K., and Nebert, D. W. (1993) The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes and nomenclature. *DNA and Cell Biol.* **12,** 1–51.
7. Sugiura, M., Sakaki, T., Yabusaki, Y., and Ohkawa, H. (1996) Cloning and expression in *Escherichia coli* and *Saccharomyces cerevisiae* of a novel tobacco cytochrome P-450-like cDNA. *Biochim. Biophys. Acta* **1308,** 231–240.
8. Scott, J. A., Collins, F. H., and Feyereisen, R. (1994) Diversity of cytochrome P450 genes in the mosquito, *Anopheles albimanus. Biochem. Biophys. Res. Commun.* **205,** 1452–1459.
9. Saitou, M. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425.
10. Henikoff, S., and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10,915–10,919.