UTSA talk April 20, 2007 short version 55 minutes

The Evolution and Genomics of Cytochrome P450

[slide 2] Cytochrome P450 is a family of enzymes of roughly 500 amino acids in size that contains a non-covalently bound heme. There is a helix-rich domain and a helix-poor domain and there is a characteristically long helix called the I-helix that contributes side chains to the oxygen binding pocket above the heme iron. These proteins are found from bacteria to mammals where they accomplish a wide diversity of reactions, mostly involving oxygen that is incorporated into a typically lipophilic substrate. P450s are not usually found in anaerobic organisms, but there are a few exceptions where P450s act as peroxygenases, probably as a protection mechanism.

The name P450 means pigment at 450 nanometers. This refers to the unusual spectrum for heme in these proteins. [slide 3] The peak position is influenced by a thiolate anion that forms the fifth ligand to the heme. If the protein is denatured the peak shifts to 420 nm and this is indicative of a dead P450 enzyme. The Cys is part of a 10 amino acid signature sequence that nearly all P450s have except those that work on peroxide substrates like allene oxide synthase from plants in the CYP74 family. This figure illustrates the structure of the signature sequence as found in CYP101A1 or P450cam (for camphor hydroxylase) This was the first bacterial P450 to be sequenced and the first P450 crystal structure to be solved. Note the close contact between the Cys and the heme iron directly beneath it.

[slide 4] The first P450 sequences were determined by protein chemistry in 1982. There was a slow accumulation of more sequences until the genome projects began to be published in 1996. The curve became steep when animals like Drosophila and C. elegans and the plant Arabidopsis were sequenced. Arabidopsis alone has 272 P450 sequences, 246 are full length presumably functional enzymes. Animal genomes have from a low of 54-57 functional P450s for dog and human to 158 for Aedes aegypti the yellow fever mosquito. I did a new gene count last week and I have named 6766 P450s.

In humans these 57 enzymes are involved in lipid and drug metabolism. They synthesize or metabolize cholesterol, steroids, prostaglandins, fatty acids, retinoic acid, bile acids, thromboxanes and leukotrienes. Drug companies are very interested in their structures for use in making better drugs [slide 5] and the newest application is pharmacogenomics and personalized medicine. This chip is approved by the FDA for genotyping 30 common polymorphisms in the CYP2D6 and CYP2C19 genes. These are major drug metabolizing P450s in humans. Knowing your genotype could prevent adverse drug reactions if you are a poor metabolizer. LabCorp does this test and they have an office here in San Antonio. CYP3A4 is not included here even though it is a very significant drug metabolizing P450 because that gene is not polymorphic so there is nothing to test for.

[slide 6] The distribution of the named P450s over the tree of life is shown here. The numbers in parentheses are the number of different P450 families. Most are from plants and animals. Note that there are very few in Archaea since most Archaea are anaerobes.

Fungi have over 1000 sequences named with many more to come from a fungal genome initiative. Fungi can have from two P450s: CYP51 and CYP61 required for making ergosterol, to over 150. Bacteria are undercounted, since I have not had time recently to name new bacterial sequences. I will come back to these at the end of my talk.

I will divide my talk into sections on animals, plants, fungi and bacteria. The first two will be longer and the last two will be brief. Since it is easier to focus on limited taxa we will start with vertebrates and then expand out to include insects and other bilateral animals like C. elegans, eventually adding in Hydra as a radial animal. Vertebrates have only 18 P450 families of the 720 families now named. They have been quite conservative over the last 425 million years since ray-finned fish diverged from tetrapods. Fish have the same 18 families as humans. Where we differ is in the subfamilies. [slide 7] A P450 family is generally composed of protein sequences that are more than 40% identical, but this has been sliding to lower values and probably should be revised to around 37%. The subfamily members are 55% identical or higher. The consequence of this is that different taxa tend to have different subfamilies. When making phylogenetic trees, it is often not possible to tell by clustering which subfamily sequences arose from the same ancestor. For example in the CYP2 family, humans have 13 subfamilies and fish have 12 subfamilies, but only two of these are shared: CYP2R1 a vitamin D 25 hydroxylase and CYP2U1 an (omega- and (omega-1)-hydroxylase of arachidonic acid. These two P450s are acting on endogenous substrates to perform a crucial conserved function. The other CYP2 subfamily relationships are unclear. [slide 8] We can use synteny to help trace back the evolutionary history of these P450s.

Synteny is the preservation of gene order among species. The Phylogenetically Inferred Groups website has a synteny viewer where you can compare two species chromosome by chromosome. [slide 9] This shows part of human chromosome 1 compared to mouse and chicken. The lines indicate orthologous genes. The mouse shares 37 genes with human in this block while the chicken has only slightly less at 32. [slide 10] When we compare human and zebrafish the synteny has fallen dramatically. Here we see only two four gene blocks and one three gene block that are conserved. [slide 11] If we now go the Ciona, a sea squirt, only two genes remain together, and I had to hunt to find a chromosome segment with this pair. So synteny is probably not useful at this time scale (about 550 million years) but it can still be sucessful with fish and human.

Let's look at some examples. [slide 12] Xenopus frogs have a large group of about 35 P450 genes in a cluster. It was not clear what these sequences matched in other species. They are flanked by the methylmalonyl CoA mutase gene and rh blood group glycoprotein. In humans a cytochrome P450 pseudogene CYP2AC1P is found between these two genes. In chicken there are two CYP2AC full length genes next to rhag. There are no P450s next to rhag in fish or Ciona and there are no clear matches by BLAST searching to P450s in fish. The CYP2AC genes may be tetrapod specific.

The next slide [slide 13] shows the human and chicken CYP2J genes in red and their neighbor HOOK1 in blue. Humans have only one CYP2J gene, while chickens have five and one pseudogene. The fish P450 cluster adjacent to Hook1 has 11 P450 genes that include four subfamilies CYP2P, 2V, 2AD and 2N. These are descended from a common

ancestor with the mammalian CYP2Js. Since humans have only the one 2J2 gene, the four fish subfamilies probably arose from a single gene after tetrapods separated from fish. The expansion of P450 gene loci is unpredictable and some times very extensive as seen in the frog 35 gene cluster.

Synteny in Fugu and zebrafish identifies the origin of the mammalian CYP2ABFGST cluster. [slide 14] The end gene CYP2S1 in humans is 10 kb from the receptor tyrosine kinase AXL which is next to TGFB1 for TGF beta. In frogs, the CYP2Q 11 gene cluster with several subfamilies shares these neighbors. In Fugu, AXL is 9 kb from CYP2Y2. In zebrafish the CYP2Y3 gene is 30 kb from TGFB1. Therefore, the fish CYP2Y genes and the frog CYP2Q cluster share an ancestor with the CYP2ABFGST cluster, once again probably a single gene.

These examples all involve subfamilies, but what about families. Unique CYP families have also emerged over time. Can synteny address the origin of a family? [slide 15] Steroids and steroid receptors do not exist in Ciona, but they are found in vertebrates. Therefore, the origin of steroid biosynthesis can be predicted to lie between Ciona and vertebrates. The CYP19 gene or aromatase is required for making estrogen from testosterone. It is seen in Amphioxus but not in Ciona. [slide 16] In humans and fish CYP19 is neighbor to gliomedin. These genes face away from each other about 13 kb apart in fish. BLAST searches of Ciona with human gliomedin find no good matches. Searches with the fish ortholog show a best hit to a sequence that is 10 kb from the Ciona CYP20 gene. These genes also face away from each other. This suggests that CYP19 arose by a gene duplication of CYP20, a gene with unknown function that is conserved all the way back to sponges. In this case the orthologous position after the duplication became CYP19 and the new location of the CYP20 gene retained the original function. If this is true CYP20 may have a substrate that is similar to testosterone. The synteny approach should also be useful in other groups like insects or plants.

[slide 17] We have been limiting oursevles to chordates, but now I would like to expand into bilateral animals. Bilateria split into two main groups the deterostomes that include chordates, Ciona and echinoderms and the protostomes that include C. elegans and Drosophila. Exanding to include bilateral animals will allow discussion of deeper time evolution in the animal P450s. Before we do that I want to illustrate the idea of P450 clans. [slide 18] When all the P450 genes from a genome are used to make a phylogenetic tree the genes naturally cluster into a few main branches. These are gene clades and they have been called clans in the P450 nomenclature. Each clan is assumed to derive from one common ancestor. All animal P450s fall into 9 clans. This figure shows 8 clans in chordates, the mitochondrial clan is left out here.

It would be desirable to determine the origin of each clan from the most recent to the most ancient, in other words to trace back the history of each P450 family to its origin. We have had a beginning with the synteny approach, but this is limited to subfamilies and maybe the CYP19 family which is in a clan of its own in chordates. One might suspect that including more diversity in the P450 sets would add more and more clans but we find that protostomes that includes the insects and worms are actually simpler. [slide 19] This tree includes C. elegans and Drosophila sequences and some related sequences from

Fugu, human and Ciona. There are only four clans, 2, 3, 4 and mito. This remains true if 350 insect P450s are included in a single tree.

The difference is due partly to loss in the protostomes. They do not have CYP51 for making sterols or the CYP20 clan. The other three missing clans appear to be deuterostome innovations. Note that the 3 and 4 clans are together. This is always true and I suspect they had a common ancestor. We will come back to this point when we talk about plants.

Now let us expand once more to include Cnidarians or radial animals. The Cnidarian group includes hydra, coral, jellyfish and sea anemones. [slide 20] This tree shows all the complete hydra P450s and some large partial CYP2 clan sequences. Some fish and insect CYPs are included. The CYP3 clan sequences are shown to emphasize that there are no CYP3 clan genes in hydra however there are CYP3 clan sequences in sea anemone, another Cnidarian. There are also no mitochondrial clan genes in hydra, but these are found in sea anemone and sponge. This suggests gene loss in Hydra. The mitochondrial P450 clan is only found in the animals. The mito P450s probably arose from a mistargeting event sending a microsomal P450 into the mitochondria where it adopted an existing electron transport system. Hydra does contain a clear CYP20 gene and an ortholog of CYP20 called CYP38 is found in sponges so CYP20 is one of the earliest animal P450s. However CYP20 is not found in plants or fungi.

These genomes suggest that the ancestor of radial animals had six P450 clans, CYP2, CYP3, CYP4, mito, CYP20 and CYP51. These were probably limited to one P450 each. Hydra lost CYP51, CYP3 and the mito clan P450. Among the bilateral animals protostomes lost CYP20 and CYP51. In deuterostomes the chordates expanded the clans by three to exploit steroids, bile acids and retinoic acid metabolism. We are waiting on a complete sponge genome and the genome of a choanoflagellate which are single celled relatives to sponges and therefore a probable sister group to all animals.

The animal P450s do so many reactions I don't have time to go into these in detail. I do want to give one interesting example in insects. The molting hormone ecydsone is synthesized by a pathway involving many hydroxylations of cholesterol by P450 enzymes. [slide 21] This shows four sequential hydroxylations carried out by four of the so-called halloween genes. Defects in the halloween genes cause proper cuticle formation to fail. These mutants are embryonic lethal. There are two additional hydroxyl groups probably added by two more halloween genes CYP307A and CYP307B but the function has not been asigned to these yet. These types of modifications are reminiscent of steroid biosynthesis in vertebrates which also employs 5-6 P450s.

That brings us to plants. [slide 22] The first plant P450 sequenced called CYP71A1 was from avocado. We now have more than 2400 named plant P450s including all of those from rice, Arabodopsis, the cottonwood tree, Physcomitrella patens (a moss) and Chlamydomonas (a green algae). Plants are much more chemically capable than animals. Their biosynthetic abilities are much greater. So, not surprisingly they have more P450s. The cottonwood tree and rice have over 300 full length P450s, many of them in large

gene clusters. Plants use these P450s to make hormones, pigments, volatiles, structural compounds like lignins in wood and cutins for waterproof barriers.

As with animals, the plant P450s sort in to a small number of clans, only 10 for seed plants, 11 if we include moss. The first four clans are shown here. [slide 23] The vertical axis is a phylogeny of plants. The top has the plant families sorted by clan. Each dot in this graph means a member of that P450 family is present in a specfic taxon. The families with the deepest presence in the phylogeny are placed to the left of each clan. Plants have 95 CYP families but if we limit ourselves to seed plants there are only 61 families. Notice that the first column is CYP51. We saw CYP51 in animals as the sterol 14 alpha demethlase required to make cholesterol. In plants it makes related phytosterols. CYP51 is also found in fungi, protists and even bacteria. It is thought to be the oldest eukaryotic P450. There is only one CYP family in the CYP51 clan.

Each plant clan is assumed to have one original sequence that gave rise to the clan. This suggests one original function that is quite ancient as in sterol biosynthesis. The CYP71 clan is the largest CYP clan in plants. The CYP71 clan is found in moss, but not algae, so it seems to be associated with arrival on land. CYP73 is the second step in the phenylpropanoid pathway that leads to lignins and flavonoids. Flavonoids are precursors to many important plant compounds such as flower pigments like anthocyanins and defense chemicals called flavonoid phytoalexins. There are over 9000 known flavonoid structures. CYP73 is a good candidate for the first CYP71 clan member. Multiple P450s including CYP98, are found later in the phenylpropanoid pathway and its offshoots.

As the sequence data comes in from EST projects and genome projects, the figure is filling in. It will be possible to say with some certainty where each family originated in the phylogeny. [slide 24] Here we have the second part of this figure. The CYP85 clan is important for gibberellin biosynthesis and brassinosteroid biosynthesis. The CYP86 clan makes cutins from fatty acids and these form the waterproof barrier or cuticle on land plants. CYP97 has three carotenoid hydroxylases that are conserved all the way back to green algae. CYP710 has recently been shown to be the C-22 sterol desaturase of plants and it is equivalent to the fungal CYP61 in the ergosterol pathway. Along with CYP51, the CYP710/CYP61 clan is also found in protozoa like trypanosomes and probably it is the second oldest eukaryotic P450.

The moss genome sheds some light on plant p450 evolution. [slide 25] There are 71 P450s in moss. The ovals mark the 11 plant clans. Comparison to the green algae P450s shows that four single family clans shown as blue ovals predate land plants. These are CYP51, CYP710, CYP97 that I just mentioned and CYP746 of unkown function. The yellow ovals are clans with multiple families that predated land plants, but the CYP families are different in algae, moss and higher plants. Finally, the green ovals are P450 clans unique to land plants. I would like to zoom in on the right side of this tree. CYP85 clan for a moment [slide 26]. The CYP711 family that we have not talked about makes a carotenoid-derived plant hormone in Arabidopsis. When CYP711 is BLASTed against animals at NCBI the best hit is always CYP5 in the CYP3 clan. When CYP5 is BLASTed against plants the best hit is always CYP711. Best reciprocal blast hits are used as evidence for orthologs. In this case CYP711 is not an ortholog of CYP5, because

CYP5 is a thromboxane A2 synthase, a vertebrate specific enzyme, but there may be orthology between the CYP711 clan and the CYP3 clan of animals. I mentioned before that the CYP3 and CYP4 clans always cluster together and probably derived from a common ancestor. When CYP4 sequences are included in phylogenetic trees of plants, they cluster with the CYP711 clan. This suggests that the animal CYP3 and CYP4 clans may have derived from a plant CYP711 ancestor. Fungi have a CYP52 clan that may also be related to CYP711.

I will give one example of what the P450s are doing in plants like I did with the ecdysone pathway in insects. The CYP85 clan includes CYP85 and CYP90. These families act in the biosynthesis of brassinosteroids, important plant growth signals. [slide 27]. The pathway is not linear but more like a grid or network. This is often the case with plant P450s. At least seven different P450s in the CYP85 clan act to synthesisze brassinolide and another from the CYP72 clan further hydroxylates it to inactivate it. This is similar to the human steroid pathway and the insect ecdysone pathway.

Let's summarize what I think are reasonable inferences about early p450 evolution [slide 28]. In this figure the ovals represent the origin of P450 clans except CYP105 and CYP55. CYP51 is considered to be first, follwed by CYP710/CYP61. These both act in the same pathway of sterol biosynthesis. Both are found as early as trypanosomes. Next came CYP711 possibly acting on a fatty acid or carotenoid substrate. The ancestors of plants, animals and fungi are predicted to have these three P450s. CYP51 has kept the same function throughout its evolution. CYP710/CYP61 also seems to keep the same funtion. This enzyme is lost in animals since we do not make a C22-23 double bond in our sterols. Green algae evolve four new clans and then land plants seperate from algae and evolve four more new clans required for coping with life on land and multicellularity, namely water barriers, structural components like lignin, hormones and protective molecules against UV light and herbivores. Vascular plants apparently lost CYP746 and the moss Physcomitrella lost CYP711, but other moss must have retained it since it is found in angiosperms. CYP55 is a rare soluble eukaryotic P450. This is explainable if it came via lateral transfer from bacteria where all P450s are soluble. CYP105 is the most likely choice for the CYP55 origin. Chlamydomonas acquired CYP55 after it evolved an intron in fungi since the two share that intron boundary. This appears to be a second lateral transfer event. Now let's take a very short look at the fungi.

[slide 29] Fungi are great heterotrophs, degrading many difficult polymers and aromatic ring compounds like lignin. In fact, if it were not for white rot fungi, the planet would be covered in non-biodegradable tree trunks. Fungi are also successful pathogens making many mycotoxins. Several fungi contain more than 150 P450 genes that are employed in these processes. Another surprising feature of fungi is they often keep genes acting in a biosynthetic pathway together as a gene cluster. This is especially true of toxin gene clusters. In synthesis of aflatoxin 25 genes invloved in that pathway are found in a single gene cluster in A.nidulans. Six P450s are required. [slide 30] These toxins are often made with polyketide synthases to form a ring structure. The ring is then decorated by P450 hydroxylations and other modifications. Note that in Nectria haematococca there are 14 PKS genes and 9 have P450s nearby. It is sometimes possible to tell what the P450 is doing by looking at its neighbors in the gene cluster. [slide 31] For example CYP65W1 is

very likely making a red pigment in Nectria since it is five genes away from a PKS that resembles a yellow pigment PKS of another fungi. Disruption of the Nectria PKS blocks red pigment formation. [slides 32-33] This slide and the next shows part of a 100 kb region with 7 P450s. There is clear homology to a phenylacetate degradation pathway in Aspergillus nidulans that involves two of these P450s and two other enzymes. It is not clear what the other 5 P450s are doing.

There is one very odd fungal group that I want to show you now. [slide 34] Remember I told you at the beginning that all P450s have a Cys thiolate anion as ligand to the heme iron. That is what gives the characteristic P450 spectrum. I found five fungal P450s that do not have the Cys. [slide 35] Here is an alignment of these P450s at the heme binding region. They are lacking the whole 10 amino acid motif. But they do have the PERF motif here and other P450 motifs like the EXXR motif. What happened to the Cys? I started looking in the sequences for any clues. [slide 36] Here is the I-helix region that usually has another motif. These sequences do not have the AGXDT motif either, but four of them have [-] Cys where the DT usually is found. [slide 37] The Thr is a residue that helps bind oxygen in the active site as it sits on the heme iron. Could the Cys at this location be serving as the thiolate anion to iron? This would turn the P450 on its head and invert it. I have two collaborator in Wales, Steve Kelly and David Lamb who are trying to express the Aspergillus fumigatus sequence and test this idea by mutating the Cys and observing the effect on the spectrum.

It is time to wrap up and I want to finish with a comment on bacterial P450s and environmental sequencing. I have 754 named bacterial P450s. There are quite a few more known from genome projects that I have not yet named, however, the largest contribution of bacterial P450s may be from J. Craig Venter and the sequencing of seawater. [slide 38] Venter's yacht the Sorcerer II has sailed around the world collecting bacteria from seawater every 200 miles as shown on this map. Just a couple of weeks ago he published the results on the first 41 sites (red dots) in Public Library of Science Biology. He found 6.1 million new protein sequences from his translated DNA. He claims this is 1.8 times the current number of all publically available protein sequences in Genbank and other sites. He also sorted these by protein family and he found 3305 new cytochrome P450 sequences, If those are added to my 6766 the new total would be over 10,000 P450s. According to this map he still has nearly 100 more sites to report with no saturation in sight; new protein families are linear with increasing sequence numbers, so that could increase the number of P450s to over 8000 new bacterial P450s from this single effort. That is a daunting prospect.

[slide 39] I put in this tree of Ernst Haeckel from 1891 to show that things have not changed much in the last hundred years. Biologists were interested in the tree of life then and they are still interested in the tree of life now [slide 40]. Only now we look at molecules not anatomy. I Leave you with the following excerpt [slide 41] taken from David Quammen's book The Reluctant Mr. Darwin. This quote was refering to a discussion on comparative genomics between the author and Dr. Futuyma.