[1] Mining Databases for Cytochrome P450 Genes

By DAVID R. NELSON

Introduction

The growth of nucleotide databases is currently exponential. GenBank release 122 (Feb. 15, 2001) had 11.7 billion base pairs (bp) of sequence. The counter on June 6, 2001 was 15.8 billion[1] for an increase of 4.1 billion bp in 111 days, or 37 million bp per day. The rate since GenBank release 123 (April 15, 2001) is 65 million bp per day. The species represented include not only human but a wide variety of prokaryotes and eukaryotes. There are now five completed eukaryote genomes,[2] not including human, and 53 complete prokaryote genomes in GenBank. More than 300 complete prokaryotic genomes are known in private databases.

This data trove offers many opportunities to find genes. The cytochrome P450s are richly represented. These genes make up 273 of about 25,000 *Arabidopsis* genes or 1% of the whole genome.[3–5] In *Drosophila* there are 90 P450 genes[6,7] in about 13,600 or 0.7% of the genome. Mice have 82 P450 genes[8] now in approximately 30,000 total genes for about 0.3%. As more and more data come in from species like zebrafish, *Xenopus,* sea urchin, and *Dictyostelium,* it becomes necessary to systematically search these sequences to find and assemble the P450 genes. The same is true for any other gene family. This chapter outlines how I have done this for some species, giving summaries of the results with emphasis on recent findings.

GenBank Substructure

GenBank is not just one database. There are multiple parts to it and when a BLAST search is done, only one section is searched at a time. Table I lists the main sections of GenBank. The default section is nr (for nonredundant), where you will search when you do a BLAST unless you actively select another section. It contains the cDNAs and genes that have been sequenced by individual laboratories over the years. Unless a submitter is sending in a batch submission of large numbers of sequences for the other sections of the database their sequence goes here. Note that nr has 3.2 billion bases or about one-fifth of the total GenBank bases. If you only search here for a match to your sequence, you are neglecting almost 80% of GenBank sequence data.

TABLE I
SIZES OF GenBank SUBDIVISIONS

Section of GenBank	Bases (millions)
nr	3214
EST	3502
Human	1631 (46%)
Mouse	734 (21%)
Others	1140 (33%)
HTGS	4356
GSS	1336
STS	89

ESTdb[9] is the division for sequencing projects generating cDNAs. Usually, large numbers of ESTs are deposited from these projects. The ESTs are broken down into human, mouse, and other ESTs to show that almost one-half are human, one-fifth are mouse, and one-third are other. Note that there are more bases in the EST section than in the nr section. The HTGS (high throughput genomic sequence[10]) section is for genomic sequencing. This is where the genome projects send their data. It is the largest section of GenBank. It has a large amount of human and mouse sequence. GSS stands for genome survey sequence.[11] Most genome projects are based on sequencing clones of differing sizes. Some are cosmids and some are bacterial artificial chromosomes (BACs). One way to tag BACs is to end sequence them to get two ~500-bp sequence markers for mapping and later assembly. The BAC end sequences go in the GSS section of GenBank. They are similar in size to ESTs but they are from genomic DNA. There are over a billion bases of the GSS sequence. Search all four sections for matches to any sequence you have, if you do not want to miss anything. There is a smaller section called STS for sequence-tagged sites[12] (only 89 million bases). These are used for mapping. Search there too, even though there is not as much data because missing exons can also be found in the STS database that are not in any other section of GenBank.

[1] www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide

[2] *Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Schizosaccharomyces pombe.*

[3] drnelson.utmem.edu/Arablinks.html

[4] www.biobase.dk/P450

[5] S. M. Paquette, S. Bak, and R. Feyereisen, *DNA Cell Biol.* **19,** 307 (2000).

[6] drnelson.utmem.edu/dros.html

[7] p450.antibes.inra.fr

[8] drnelson.utmem.edu/fasta.mouse.P450s.html

[9] www.ncbi.nlm.nih.gov/dbEST/index.html

[10] www.ncbi.nlm.nih.gov/HTGS

[11] www.ncbi.nlm.nih.gov/dbGSS/index.html

[12] www.ncbi.nlm.nih.gov/dbSTS/index.html

TABLE II
MAJOR DIVISIONS OF nr DATABASE[a]

nr Section	Entries	Bases (millions)
PRI (primate)	162,557	1512
ROD (rodent)	65,765	124
MAM (other mammal)	30,048	26
VRT (other vertebrate)	54,414	49
INV (invertebrate)	87,610	431
PLN (not animal, bacterial, or viral)	153,924	381
BCT (bacteria, archaea)	104,940	283
VRL (viral)	118,665	104
Total	777,723	2910

[a] GenBank Release 123, April 15, 2001, nr Database Sections.

Looking more closely at the nr section, release notes for the April 15 release 123 show the following.[13] The GenBank release is broken into several hundred files. If the EST, GSS, STS, and HTGS files are ignored, nr is what is left. The nr sequences are categorized into eight groups listed in Table II. These designations appear in the top line of any sequence from the nr section of GenBank. Primate is the largest section with the human sequence in it. Rodent is about 10 times smaller, with other mammal and other vertebrate (zebrafish, *Xenopus,* etc.) tiny by comparison. Invertebrate is fairly large, as it has both fly and *Caenorhabditis elegans* sequences in it. Plant is a catch-all section that has every species that is not animal, bacterial, or viral. This includes fungi and protists, even though they are not plants. It is fairly large because *Arabidopsis* is included here. Bacteria is also fairly large, as all 53 public prokaryotic genomes are here.[14] Viral rounds out the nr section. There are 607 complete viral genomes in GenBank.

When doing a BLAST search,[15] you can specify any of these categories except invertebrate. You can also select all animals (metazoa) or all arthropods, which gives you a subset of invertebrates, but not all. There are also 27 common species in the BLAST pull-down menu; however, you can type in any taxon name in another window to limit a search. For instance, if you wanted to search only soybean sequences, you could enter *Glycine max*. It is often important to do this to reduce noise. The BLAST output can limit your results to 100 alignments. In a sample search of the db EST with a known *Arabidopsis* sequence with the filter set for Viridiplantae (green plants), 100 alignments were recovered but only 26 were from soybean. When the filter was set for soybean, 100 hits recovered were all soybean. The filter eliminated the noise from all the other plant sequences.

TABLE III
dbEST RELEASE 051101 SUMMARY BY ORGANISM[a]

Organism	Sequences
Homo sapiens (human)	3,510,107
Mus musculus + *domesticus* (mouse)	1,961,530
Rattus sp. (rat)	285,326
Bos taurus (cattle)	168,677
Glycine max (soybean)	168,157
Drosophila melanogaster (fruit fly)	125,727
Medicago truncatula (barrel medic)	122,365
Lycopersicon esculentum (tomato)	115,070
Arabidopsis thaliana (thale cress)	113,000
Caenorhabditis elegans (nematode)	109,215
Xenopus laevis (African clawed frog)	91,558
Zea mays (maize)	89,125
Danio rerio (zebrafish)	88,276
Oryza sativa (rice)	75,057
Hordeum vulgare (barley)	68,480
Sorghum bicolor (sorghum)	65,092
Chlamydomonas reinhardtii (green algae)	64,973
Sus scrofa (pig)	64,709
Triticum aestivum (wheat)	60,022

[a] Number of public entries: 7,965,906. May 11, 2001.

The db EST has a summary by organism that shows the most sequenced species.[16] The top 19 organisms are shown in Table III. Human and mouse are at the top with more than 5 million EST sequences between them. The top 4 are mammals, but 9 out of the top 19 are plants. Surprisingly, cattle have been sampled pretty heavily. The human CYP26C1 gene was missing exons 2 and 3 from all human sources I could find, including Celera Genomics public data. An exon 1 and 2 fragment is present in cattle ESTs. I know this because cattle exon 1 is 91% identical to the human sequence. In May 2001, the two missing human exons were deposited in a new version of a human genomic sequence so it pays to look often for new hits. The chance of finding a gene or a close homolog in the EST database is good, even rare transcripts may be represented. For humans, with approximately 30,000 genes, there are on average over 100 ESTs per gene. This does not mean every gene is represented. There are five human P450s that are not pseudogenes yet they have no ESTs associated with them. These are CYP2A13, 2C19, 4A22, 26C1, and 27C1. This shows that the EST database does not have every gene represented

[13] ftp.ncbi.nih.gov/genbank/gbrel.txt

[14] www.ncbi.nlm.nih.gov/PMGifs/Genomes/a_g.html and http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/eub_g.html

[15] www.ncbi.nlm.nih.gov/blast

[16] www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

TABLE IV
dbGSS RELEASE 051101 SUMMARY BY ORGANISM[a]

Organism	Sequences
Homo sapiens (human)	869,266
Mus musculus (mouse)	835,820
Tetraodon nigroviridis (freshwater pufferfish)	188,963
Oryza sativa (rice)	93,107
Trypanosoma brucei (African sleeping sickness)	90,540
Strongylocentrotus purpuratus (sea urchin)	76,019
Arabidopsis thaliana (thale cress)	61,265
Entamoeba histolytica (human parasite)	49,129
Drosophila melanogaster (fruit fly)	45,143
Fugu rubripes (Japanese pufferfish)	42,896
Magnaporthe grisea (rice blast fungus)	12,674
Trypanosoma cruzi (Chagas' disease)	12,232
Leishmania major (Leishmaniasis)	11,929
Plasmodium vivax (malaria)	10,682
Saccharomyces cerevisiae (Baker's yeast)	7,637
Anopheles gambiae (malaria mosquito)	7,514
Cryptosporidium parvum (human parasite)	7,082
Plasmodium berghei (rodent malaria)	5,476
Glycine max (soybean)	4,818

[a] Number of public entries: 2,493,887. May 11, 2001.

even though there are 3.5 million sequences. These five P450s represent 9% of the known human P450s so gene coverage in the human EST database is about 91% based on known P450 genes. There still may be ESTs from the 3 prime (3′) end of the message, but it is difficult to recognize them without a 3′-untranslated sequence for comparison.

The GSSdb has over 1.3 billion bases of genomic sequences from many different organisms.[17] The top 19 are shown in Table IV. This section of the database is especially interesting because it has a large number of fish sequences from both freshwater and salt water pufferfish. There are even 76,000 GSSs from sea urchin. There are also many parasite ESTs here, including *Plasmodium* (malaria), *Leishmania, Entamoeba, Trypanosoma,* and *Cryptosporidium*. The vertebrate sequences can help in assembling human genes by clarifying the positions of introns. It can be quite difficult to identify intron boundaries when you only have a DNA sequence of a single species, but it is much easier with a human–mouse or human–fish pair. The GSSdb is a very valuable resource.

The last section is the HTGS section.[10] The HTGS section has sequence that is not finished. Once an HTGS sequence is finished it gets annotated and moved to the

[17] www.ncbi.nlm.nih.gov/dbGSS/dbGSS_summary.html

TABLE V
NUMBER OF HIGH THROUGHPUT GENOMIC SEQUENCE ENTRIES[a]

Organism	Finished sequence (nr)	Organism	Working draft (HTGS section)
Human	11,894	*Giardia*	53,667
C. elegans	2,363	Human	24,442
Drosophila	1,881	*Drosophila*	5,165
Arabidopsis	769	Mouse	1,909
Mouse	231	*Rice*	391
Rice	154	*C. elegans*	201
Other	142	Other	496
Total	17,435		86,309

[a] In GenBank (5/20/01).

nr database so HTG sequences can be found in both nr and HTGS. Table V shows the top six species for HTG sequences either in finished form or working draft sequences. The top four in the finished category are from model organisms that are completely sequenced or human that is mostly sequenced. On the working draft side there is a surprise in *Giardia* with 53,000 entries. *Giardia* is the parasite that causes hiker's diarrhea and these reads are short reads of about 500–800 bp. The human and mouse entries are much larger, being up to 200,000 bp and including many fragments. Do not be fooled by large numbers of entries, it is base pairs that really count. *Giardia* is anaerobic and does not appear to contain any P450 genes.

Other Databases and Systematic Search Strategy

GenBank is not the only place to go for sequence data. There are many sequencing facilities around the world that have searchable databases for specific organisms. A prime example is the *Dictyostelium discoideum* genome project. This project is being carried out by multiple laboratories that have set up one common BLAST server for sequence data.[18] Most of this data is not in GenBank. The strategy for finding P450s in this data is a general strategy that can be applied to other species, with either genomic or cDNA sequences or preferably both. The initial step is to do BLAST searches with a wide variety of sequences that are appropriate. For plants, searches with one member of each plant family would be desired, but that would be over 50 searches, so this might be shortened by selecting one member from each plant clan. For vertebrates, one member from each mammalian family or 18 sequences would be recommended. *Dictyostelium* was searched with the 18 mammalian sequences to find all possible hits. For searches done at GenBank, the species should be selected and the HTGS, EST, nr GSS, and STS sections should be searched.

[18] www.sdsc.edu/mpr/dicty

After the first search, the output needs to be examined and the sequence fragments that are true P450s need to be collected in a file. I make two files for this data: one is just a copy of the blast output, which eventually gets sorted into sequence bins and the other file is a FASTA format file of just the sequence fragments by themselves with an identifier line at the top of each sequence that begins with >. Each sequence is given an arbitrary number, 1, 2, 3, etc., to identify that sequence bin. A third file is made that has the accession numbers listed in alphabetical order, followed by the arbitrary sequence number. Once the FASTA file is made from the first search, it is used at the Proweb.org website as a database file for blast searching.[19] Here, all the fragments are searched against this file one by one to identify overlapping fragments. These are combined. Their sequence numbers are also combined, so if sequences 8 and 12 are really the same protein, that sequence bin becomes 8, 12 and the sequences are joined in the FASTA file to make a longer sequence. This process builds up contigs. In a species like human with many named P450 genes, a FASTA file can be used that has all the known human sequences present so any blast hits can be identified quickly.

The second BLAST search output is compared with the accession number list from the first search. Any new accession numbers that are real P450 sequences get added to the master accession number list. The sequences get added to the FASTA file and checked for any overlaps as before. If the sequence is new, it gets a new bin number. If the sequence is a match, it gets combined in an existing bin. The first three to five searches usually identify new sequence bins, but after that the new sequences become quite rare. It is unusual to find new sequence hits after seven or eight rounds of this process. After all the test sequences have been searched and the contigs sorted into the smallest number of bins, it is time to try to finish out the individual contigs. This process is different for EST and genomic data.

ESTs can be extended as far as the ends of the sequence will permit. The blast search output probably stopped before reaching the end of coding sequence in the EST so it is necessary to go get the EST and translate it in the three frames of the correct strand. Two useful sites for this are the JustBio.com[20] translator or the EMBOSS translator[21] at EBI. The translations need to be examined for possible extensions of the sequence out to start methionines or stop codons. Frameshifts should be identified by comparison to the most similar reference sequence, which may be from the species being searched. The result may not be completely correct, but errors will correct themselves later when more accurate sequence is available. Extensions should be checked against the FASTA file to see if any contigs can be joined. The names of sequences may be helpful in constructing contigs. Often, clone names indicate whether they are sequenced from the 5′ or 3′ end of a clone. Clones with the same prefix come from the same clone and these are probably from opposite ends of the same gene. The examples JC1b30e03.s1 and JC1b30e03.r1 are standard and reverse reads from the same *Dictyostelium* clone and from the same gene. The set of alphabetized accession numbers should be examined for these clone pairs and the sequence bins should probably be combined. The only problem that may occur in doing this is due to gene clusters in genomic DNA or chimeras in EST data. These will probably resolve themselves later.

Genomic data should be fetched and the closest reference sequence should be used in a TBLASTN search against just that one DNA sequence using the Proweb.org server[19] rather than the whole database. This may reveal weak matches not reported earlier for the middle, less conserved part of the gene. This will also reveal gene clusters by giving multiple alignments to the same region of the reference sequence. These should have been detected before. The middle of a P450 is the hardest to find. These regions can be found in genomic DNA by locating the conserved C-helix region or N-terminal proline-rich region and the I-helix region that are easier to recognize. Once the middle region is bracketed by these sequence blocks, it is possible to translate the region in between in three frames and look for possible exons. It is very helpful to have a complete gene sequence with known intron–exon boundaries from the same species to help in this process. The exons are mostly conserved in size and phase in a given family, so looking for the GT–AG pairs is easier. At this point, EST data are very useful to spot these boundaries, even if the EST is not a 100% match. If a complete gene can be assembled, the FASTA file should be searched again with the new sequence to try to join contigs. This process is an iterative one.

Dictyostelium P450s have been sorted into 55 contig bins with 705 accession numbers.[22] There are a minimum of 43 P450s in *Dictyostelium* based on overlapping C-terminal sequence regions. Twenty six P450s are complete sequences. An alignment of these sequences is posted.[23] The sequences fall into 15 families CYP51, 508, 513–525. Currently, there are 12 fragments not named because they are too short. There are eight pseudogenes. CYP51 is the only family shared with plant and animal species; however, CYP524 shows 34% identity to CYP710A1 and A2. It is possible that the common ancestor with plants had two P450s, a CYP51 and CYP710/CYP524 ancestor.

The Human Genome

This strategy has been applied to the human HTGS section of GenBank. It has not been carried out on nr, EST, or GSS sections. With over 3.5 million ESTs for human, it is expected that there will be thousands of accession numbers for

[19] www.proweb.org/proweb/Tools/WU-blast.html

[20] www.justbio.com

[21] www.ebi.ac.uk/emboss/transeq

[22] drnelson.utmem.edu/LowerEuks.html

[23] drnelson.utmem.edu/dicty54aln.html

ESTs of P450s. A similarly high number will exist in nr because of duplicate entries and multiple accessions for exons. The result from the HTGS section done in September 2000 was 180 sequence entries on 130 accession numbers. Some accession numbers have more than one entry because they contain more than one gene. For example, AC020705 has CYP1A1 and 1A2. The results of this search have been sorted by CYP name,[24] chromosome,[25] and accession number.[26] There are 57 human P450s that are not pseudogenes.[27] The number of pseudogenes is not determined yet and it is going to be difficult to decide what constitutes a pseudogene. Some lone exons are found outside complete genes. Are they pseudogenes or part of the intact gene? There may be a need for some new terms to describe the detritus surrounding human genes.

The count of human P450s has increased by seven in the past 18 months. The new genes are 27C1 (April 25, 2000), 2U1 (April 25, 2000), 4V2 (April 25, 2000), 26C1 (July 19, 2000), 2W1 (Sept. 26, 2000), 20 (March 9, 2001), and 4A22 (May 23, 2001). The dates given are the dates I became aware of the sequences, not the first appearance in the database. Since the last systematic hunt for new human P450 genes (September 2000), only two new genes have been found. CYP4A22 is 95% identical to CYP4A11 (see alignment[28]) and there is some question whether it is an allele of 4A11. All mRNA sequences match 4A11, but three independent genomic sequences match 4A22 (AL390073, AF208532,[29] AL135960). The genomic sequence that matches the 4A11 mRNA is not available yet. CYP20 was identified in GenBank as a possible new P450 from human pheochromocytoma in September 2000 (AF183412). It did not turn up in the searches because it is a poor match to existing families (only 23% to CYP3A4 over 437 amino acids). The heme signature is short by one amino acid, but it does have the PERF motif and the EXXR motif. It does not have the typical AGX(D,E)T sequence at the I-helix oxygen-binding region, so the substrate for this P450 may have oxygen included as in the allene oxide synthase. I actually found this sequence when I did a text search for the term P450 in new GenBank entries. More data including sequences and accession numbers are available at the P450 home page.

One significant outcome of the human genome project has been the completion of known P450 fragments. At present, all of the 57 known human P450s are complete in their coding regions. This is a very recent occurrence, with the last missing pieces being added to the database on May 2, 2001 (AL358613.11 finished CYP26C1). Only one fragment (exon 5 of CYP27C1) is known solely based

[24] drnelson.utmem.edu/hum.htgs.byname.pdf

[25] drnelson.utmem.edu/hum.htgs.bychr.pdf

[26] drnelson.utmem.edu/hum.htgs.pdf

[27] drnelson.utmem.edu/hum.html

[28] drnelson.utmem.edu/4a22and4a11.html

[29] H. Kawashima, T. Naganuma, E. Kusunose, T. Kono, R. Yasumoto, K. Sugimura, and T. Kishimoto, *Arch. Biochem. Biophys.* **378,** 333 (2000).

on Celera Genomics public data. The rest are found in public project data. This does not mean that all genes are complete in a single contig. Many are still split in numerous unjoined DNA sequences. The long wait to complete the CYP26C1 and CYP27C1 genes was exacerbated by the fact that neither of these genes had any human or mouse ESTs in the db EST. The fragmentary structure and lack of ESTs might suggest that they are pseudogenes, but that possibility has been eliminated by finding highly similar EST sequences from bovine for CYP26C1 (95%) and from *Xenopus* (AW637606, BE669236 74%), bovine (BE723057 89%), and chicken (BE139968 71%) for 27C1. The absence of mouse or human ESTs from among more than five million mouse and human ESTs in the database suggests that these genes are expressed in a limited manner. They might be developmentally significant.

Mouse P450s

Even though the mouse genome project is not as far along as the human project, there are more P450 sequences known from the mouse. The most recent compilation shows 85 mouse P450 genes[8] and only three of these are pseudogenes. The main compilation was done April 19, 2001 with new genes being added up to May 17, 2002. There is still one human gene without a mouse ortholog discovered yet, but this is expected to exist (CYP27C1) so the mouse total should be 83 genes plus the three pseudogenes. That is 26 more than seen in human. One of the curious facts about the mouse is the presence of intact genes for Cyp2g1 and Cyp2t4. These are only found as pseudogenes in humans. The mouse also has nine intact Cyp4f genes and one pseudogene (Cyp4f13–18, 4f28, 4f29p, 4f30, 4f31), whereas humans have six intact 4fs and eight pseudogenes (CYP4F2, 3, 8, 9P, 10P, 11, 12, 22, 23P–28P). This may be due to the incomplete nature of the mouse genome sequence. More pseudogenes will probably be found as the mouse project matures. Subfamilies in the mouse that have more than a single member are usually larger than the same subfamily in humans. The 2C subfamily in humans has only four genes but mouse has eight plus two pseudogenes. The expansion of P450 subfamilies in the mouse is probably due to a more varied diet and a greater dependence on olfaction.

Fish, *Xenopus*, and Sea Urchin

Evolution of the P450 families in vertebrates can only be determined by sampling outside the mammals in an evolutionary spectrum. This is being done for several fish species, including zebrafish, *Tetraodon nigroviridis* (freshwater pufferfish), and *Takifugu rubripes* (Japanese pufferfish). The Fugu project was advertised as generating 3× coverage and about 95% of the genome represented by

April 2001.[30] The Fugu BLAST server[31] had 222 million bases of sequence in 412,012 sequences on May 31, 2001. Only 189,000 GSS sequences are in GenBank, so this is another example of a non-GenBank sequence database. As a test case, a BLAST with 27C1 found the first three exons of the Fugu 27C1 sequence. A systematic search of this data will provide a valuable set of P450s from a nonmammalian vertebrate source. In the meantime, two searches have been done on smaller data sets that are of similar interest. *Xenopus* EST data have been searched exhaustively for all P450 hits as described earlier for *Dictyostelium.* Searches of 91,558 *Xenopus* ESTs resulted in finding 120 with the P450 sequence. These were sorted into 44 contigs.[32] There are at least 27 sequences represented from 11 of the 18 mammalian families. Seven protein sequences are complete (1A6, 1A7, 2Q1, 4T4, 17, 19, 26A). Three others are nearly complete (4T3, 27A, 51). Nineteen contigs belong to the CYP2 family, including a clear CYP2R1 ortholog. Based on sequence alignments, there are a minimum of 11 different CYP2 P450s. One contig is in the CYP3 family. Seven contigs are in the CYP4 family. There are at least six members in the *Xenopus* CYP4 family Two fragments are in the CYP8B subfamily and two more are in the CYP46 family. The 26 family has three different sequences corresponding to orthologs of 26A, 26B, and 26C of mammals. There is a clear 27C1 ortholog. Sequences are absent so far from CYP5, CYP7, CYP11, CYP20, CYP21, CYP24, and CYP39. This is a good sampling from only 92,000 ESTs.

The purple sea urchin *Strongylocentrotus purpuratus* has 76,012 GSS fragments available at GenBank. These have been searched for P450s, and 51 sequences[33] were found to contain P450 genomic DNA. The sequence similarity drops compared to fish and *Xenopus* sequences and it is harder to assign sequences to specific families and subfamilies. There are sequences that show good identity to CYP2D26, CYP2U1, and CYP4V3 sequences. Most other fragments are less similar. The problem with GSS fragments is their genomic nature. The introns in sea urchin are large enough that the GSS fragments only contain one exon or part of one exon each and these cannot be joined by overlap. Many more sequences will be needed along with ESTs or close homologues in other species to assemble these into whole genes.

Arabidopsis and Rice

The plant kingdom has many more P450 genes per species than animals. We know that from the complete set of 273 P450 genes[3–5] from *Arabidopsis*. The next plant to be completely sequenced will be rice, probably in 2002. It is desirable to compare these monocot and eudicot genomes and see what the difference is between their P450s. This project has been started, but it is awaiting access to Monsanto 5× sequence data before it can be pushed forward. In September of 2000, BLAST searches were performed on the nr, EST, HTGS, and GSS sections of GenBank using one member from each P450 clan in plants (9 different sequences). After these 36 searches were done, 352 accession numbers of P450 sequences had been identified. These contained 389 P450 sequence fragments, as some accession numbers held clusters of 9 or even 10 P450 genes. All sequences from the BLAST output were compared against each other. The result was 209 contigs, of which 40 were full-length P450 sequences, and 169 were partial P450s. All 209 sequences were compared to a database of all *Arabidopsis* P450s plus five additional P450s from families not present in *Arabidopsis* (CYP80, CYP92, CYP99, CYP719, CYP723). This identified the fragments to specific families. Once this had been done, there were 20/50 plant P450 families that had no P450s present in the rice set. To be sure that no small fragments in the EST or GSS section of GenBank had been overlooked, especially fragments from less conserved regions, such as the extreme C-terminal, BLAST searches were done with one member from each of these 20 families against the EST or GSS section of GenBank limited to rice. This identified six additional entries in three different families. Three of these were identical sequences containing a small fragment upstream of the I-helix in CYP73. One was from a different region of CYP73. One was a single small exon including the heme-binding site of a CYP708 P450. One was from a CYP92 sequence. This increased the accession number count to 358 and the contig count to 213. Comparing families between rice and *Arabidopsis* genomes, 32 of 45 families (71%) were present in both species. CYP92 was found in rice but not in *Arabidopsis*. Because the rice genome is far from complete in public databases, it appears that most plant P450 families existed before the monocot–dicot divergence. As more data come in, the number of families missing between the two species will probably drop to a very small number. These may be specific to eudicots or even of more limited range. As mentioned earlier, five families are not seen in *Arabidopsis,* and this specialization may occur in each major lineage of plants. The detailed results of these searches are posted.[34]

Conclusions

Data from which P450 protein sequences can be derived are growing at a pace that is faster than it can be effectively searched. The process of systematically finding and assembling all P450 sequences for a given species is slow and tedious. I have posed this problem to computer scientists at The Tullahoma Space Institute

[30] D. Cyranoski and P. Smaglik, *Nature* **408,** 6 (2000).

[31] bahama.jgi-psf.org/prod/bin/blast_fugu.cgi

[32] drnelson.utmem.edu/Xenopus.seqs.html

[33] drnelson.utmem.edu/seaurchin.html

[34] drnelson.utmem.edu/rice.report.html

and Oak Ridge National Laboratories in Tennessee. They expressed interest in automating the steps of the procedure, but assumed that it would take on the form of an expert system, as some judgment is required during the process. Perhaps some of the more repetitive steps can be automated, saving time for people to make the harder decisions. This will be required to move forward in genome annotation because there are not enough people to dedicate themselves to this process for each protein family.

The P450 family has reached complete status in several eukaryotic species. Results from *Arabidopsis, Dictyostelium,* and three animals suggest that the evolutionary history of P450s in the crown group of eukaryotes is largely independent. The main lines starting from a billion or more years ago have been diverging from very few (two or three) P450 genes shared in the crown group common ancestor. The finding that diatoms have a CYP97 in common with plants points to CYP51 and CYP97 being in the last ancestor of plants and Stramenopiles. CYP524 in *Dictyostelium* may have a common ancestor with CYP710 in plants. Aside from these similarities across kingdoms, most P450s in different kingdoms seem to be derived independently. More complete data sets will clarify this and identify the precursor P450s present in the earliest eukaryotes. Deeper sampling of animals such as the sea urchin, *Ciona intestinalis* (the most primitive creature known with a notochord), proposed by the Department of Energy,[35] a radial animal like a jellyfish, and a sponge would make the evolutionary history of P450s (and our whole genomes) in animals apparent. I look forward to the accumulation of data and to the accumulation of better tools to analyze data.

[35] www.er.doe.gov/production/ober/berac/genome-to-life-rpt.html