1. Specific Aims and Deliverables

Improvements in sequencing technologies have resulted in the elucidation of genome content in recent years. There are now hundreds of prokaryotic and eukaryotic organisms for which the genome or the majority of the gene complement has been sequenced[1-3]. However, only a handful of these organisms have a large body of literature and corresponding community-level databases. There is a growing need to place the sequenced and annotated genomes in a biochemical context in order to facilitate discovery of enzymes and engineering metabolism. We propose to create a network of plant metabolism databases by leveraging our expertise in annotating genomes, generating metabolic pathway databases, curating biochemical information from the literature, and forming extensive network of collaborations with biological databases and biochemistry researchers. This project has **four specific aims** and will provide **four major deliverables** to the plant and bioinformatic research communities. The **first aim** is to create an integrated, non-redundant, and comprehensive plant metabolic pathway database, PlantCyc, which will be actively curated from information in the literature. PlantCyc will contain both experimentally verified and hypothetical (either based on paper biochemistry or computationally predicted) pathway and enzyme information, with tagging of all evidence as experimental or computational. The **second aim** is to create PGDBs for 23 plant species whose genomes have been sequenced or have a large body of transcript sequences. We will develop a set of standard methods to improve the accuracy and sensitivity of metabolic content predictions for plants and use these methods, in addition to the existing prediction software[4], to generate the databases. The **third aim** is to improve the content of PlantCyc and the PGDBs. Each of the newly created PGDBs will be validated, imported into PlantCyc and extensively curated with experimentally verified data from the literature. The **fourth aim** is to create an alliance of plant metabolic pathway database developers and users to minimize duplication of efforts and cost while maximizing the breadth and depth of curated data content. In addition, this aim will broaden the participation of diverse group of biochemists, database curators, and genomic/bioinformatic researchers as well as train wet lab scientists in the use of bioinformatic and database tools and methods. This project will generate **four major deliverables**: 1) PlantCyc; 2) 23 species-specific PGDBs; 3) consistently annotated gene sequences encoding enzymes; and 4) training materials. The databases will be available via a web site for searching, browsing and data analysis, via local installation on a user's computer under an open database license, as bulk data files in different formats including tab-delimited text files, SBML XML files, MOL files, and Cytoscape-formatted files. The training materials will be available as HTML pages via the website, PDF and PPT files.

2. Relevance and Rationale

The past two decades have seen the advent of plant-based technologies. Developing new technologies using plants continues to gain momentum because of the ever-expanding demand for production of biofuel, food, medicine and animal feed from plants. The long-term goal of generating those technologies has prompted the sequencing of genomes and genes. To date, genome and mRNA sequences are available for a number of representative plant species such as *Arabidopsis* (dicot), rice (monocot) and poplar (tree), which have now all been fully sequenced[5-8]. More recently, additional genome sequencing projects have flourished thanks to the continued improvements to sequencing technology[9]. Currently, 86 plant genome projects are registered with NCBI, 25 of which are underway and 6 of which are completed. In addition, over 10 million mRNA sequences (as Expressed Sequence Tags (ESTs) or cDNAs) can be found in NCBI under '*viridiplantae*' (green plants). These sequences represent over 130 plant species for which large-scale EST or cDNA collections are publicly available for functional annotation.

There is a growing need to place the sequenced and annotated genomes in a biochemical context in order to facilitate discovery of enzymes and engineering of metabolism. Several databases contain metabolism information for a wide range of organisms such as the Kyoto

Encyclopedia of Genes and Genomes (KEGG)[10-12] Enzymes and Metabolic Pathways (EMP)[13], and MetaCyc[14]. Each has its strengths and weaknesses, some of which have been reviewed[11, 15]. As they currently stand, their usefulness as reference databases for plant genomes is somewhat limited for one or more of the following reasons: 1) pathways are not associated with literature citations impeding accuracy assessments; 2) pathways tend to be composites made up of variants from many different species and are therefore not accurate for any single species; 3) they include relatively few pathways specific to plants.

The authors of this proposal have significantly improved the plant content of the MetaCyc database over the past few years[16]. The number of plant pathways increased to 266 from 18. While MetaCyc now contains a significant number of plant pathways, it is far from being comprehensive in the coverage of plant metabolism. First, it is missing a large number of plant pathways published in literature and the current mode of in-house curation is not sufficient to bring in all of the experimentally studied pathways. Second, it does not contain all of the plant sequences that are annotated as enzymes. Third, it does not include published putative pathways that present the most likely metabolic sequence based on preliminary tracer labeling and 'paper chemistry'.

The Pathway Tools software[4] in conjunction with a reference database such as MetaCyc, can be used as a reference database to create new pathway genome databases (PGDBs) from annotated genes. The Pathway Tools software also supports manual curation, web publication, analysis, visualization, and data sharing of the derived databases. MetaCyc was initially originated from EcoCyc (a PGDB for *E. coli*) and the majority of the pathways in MetaCyc are microbial in origin. Therefore, MetaCyc is a good reference database for predicting prokaryotic PGDBs and over 200 have been generated[17]. However, most of these PGDBs have not yet been curated and likely contain many false positives. Currently there is no established infrastructure to curate these PGDBs. The situation is worse for plants and other eukaryotes. Only three species-specific metabolism databases for plants are publicly available: SoyBase[18], AraCyc[19], and RiceCyc[20]. These databases vary in their level of curation. Each group has a minimal number of curators (0.5 to 1 FTE) and therefore cannot easily incorporate large amounts of data from the literature. A more integrated network to facilitate shared curation of plant metabolic pathways is desperately needed.

A comprehensive set of plant pathway databases with integrated gene, enzyme, reaction and metabolite information is an excellent platform for many aspects of basic research and its applications: 1) it depicts the biochemical components of an organism; 2) it aids in comparative studies of pathways across species to facilitate metabolic engineering for the improvement of crops; 3) it can be used as a platform to integrate and analyze data from large-scale experiments such as gene expression, protein expression, or metabolite profiling; finally 4) by facilitating identification of pathway steps lacking assigned genes, or having genes assigned solely by computational prediction, it can be used to highlight the biochemical steps for which genes have yet to be identified and experimentally characterized. The user base of pathway databases has been growing steadily. For example, AraCyc and MetaCyc serve 84,835 and 33,012 MetaCyc page views via web accesses per month (May 2006), respectively. In addition, a total of 58 groups in 20 countries have downloaded the AraCyc database in the first 9 months since it became an open database in October 2005. Over 20% of the groups are from developing countries. MetaCyc, in conjunction with the Pathway Tools software, has been licensed to 1700 groups worldwide since 1999. From the results of a user survey we conducted with the 58 AraCyc licensees, we see a high level of appreciation for the AraCyc pathway resource we provide. Some use AraCyc as a learning or teaching tool, others use it as a data mining tool for functional genomics results (e.g. microarray), or for metabolic modeling. The vast majority of the survey responses also indicated great interest in conducting comparative pathway analysis across species and thus expressed a strong desire for a plant metabolism database that includes many plant species. Although genome or EST sequences data have become available for many other plant species, only a few of the organisms have been placed into a metabolic pathway context, largely due to lack of the infrastructure to annotate and curate species-specific PGDBs.

3. Results from Prior NSF Support

Prior projects most relevant for this proposal have been supported by a grant from NIH for MetaCyc, a donation from Pioneer-HiBred for AraCyc, and a grant from NSF to support TAIR. The NSF grant for TAIR is entitled "TAIR: The Arabidopsis Information Resource" (Award number: 0417062; Duration: 5 years; Award Amount: $7,988,952) and supports 0.5 FTE curator for work on AraCyc.

Summary of Results from NSF grant (The Arabidopsis Information Resource)
During our initial award period, we developed TAIR to serve as a comprehensive Web-based, public and user-friendly information resource for the model plant *Arabidopsis thaliana,* which contains all available genomic and genetic data.

Database infrastructure and usage statistics: Since its inception in September 1999, TAIR has grown dramatically in usage, impact, and quantity, quality and diversity of data. There are 16,005 registered users and 6,103 laboratories with TAIR. With the delivery of approximately 1 million page views to 30,000 unique IP addresses via the Web per month, TAIR is one of the largest organism-based biological research community databases. The TAIR system is built on industry-standard relational database software (Sybase) and data access software written in Java using Servlets and Java Server Pages (JSP) technologies. The database contains major data types that range from genome annotations, clones, sequences, germplasms, and polymorphisms and represents the most complete knowledge-base available for Arabidopsis thaliana. In addition to the basic data access tools, we developed several visualization and analysis tools for browsing and manipulating data[21, 22].

Data curation We maintain and update the function and structure of all Arabidopsis genes. We set up a comprehensive software and analysis environment to support the genome structural annotation efforts, which includes a curation database and API called ADB, a computational gene structure update tool from TIGR (PASA), a manual annotation tool from Flybase (Apollo), sequence analysis pipelines for mapping new sequences such as cDNAs, ESTs and T-DNA insertions onto the genome, and XML-based data pipelines for exchanging data between the ADB database, NCBI and TAIR. TAIR released its first version of the Arabidopsis genome annotation on November 11, 2005. This release contains a total of 26,751 protein-coding genes, 3,818 pseudogenes and 838 non-coding RNA genes (31,407 genes in all), providing the most comprehensive Arabidopsis genome annotation to date. Since the first TAIR genome release in November 2005 we have been making improvements to our suite of genome annotation tools to improve the quality of automated gene structure updates and provide more automated quality control checks on gene structure data. Similarly, we have made substantial progress in both manual and automated functional annotation of the Arabidopsis genome. We have been using controlled vocabularies (CV) for gene annotations (Gene Ontology (GO) terms[23], as well as Plant Ontology (PO) terms[24]). These terms have been used to describe the subcellular location of a gene product, its function, the biological processes it is involved in, and the anatomical parts and developmental stages in which the gene is expressed. The reliability of the CV term/gene assignments is given by associating evidence codes and evidence descriptions, all supported by traceable references. More than 7000 Arabidopsis genes now have annotations from the literature describing their biological roles, biochemical activities, subcellular location and mRNA expression patterns. In addition to our manual literature curation, we used computational methods to assign CV terms to the whole genome complement, including many genes that have not been described in the literature. Currently all of the protein and RNA-coding genes in TAIR have at least one GO annotation.

Collaborations To enhance data access for the research community the TAIR project has established a large number of collaborations with researchers, other biological databases, and commercial organizations. One of the most significant of these was the incorporation of database functions for the Arabidopsis Biological Resource Center (ABRC)[25] into TAIR. We also collaborated with Cereon Genomics, a subsidiary of Monsanto, to release over 80 Mb of *Landsberg erecta* sequence and over 57,000 SNPs to the public[26]. To our knowledge, this was the first successful collaboration between the private sector and a publicly funded resource to make privately funded data freely available to the public. We have been involved in extensive collaborations with other biological databases such as Gramene[27, 28], MaizeGDB[29], SGD[30], MGD[31], and Flybase[32] and established community standards in controlled vocabulary development and software.

Summary of Preliminary Results from MetaCyc and AraCyc

During our MetaCyc award period, we have significantly enhanced the plant content of MetaCyc by curating over 250 experimentally determined pathways in plants. We also significantly improved ontologies used by MetaCyc and all other Pathway/Genome DataBases (PGDBs) derived from MetaCyc. In year 2001, we created AraCyc, a PGDB for *Arabidopsis thaliana*, using the Pathway Tools software. Since its inception, the quality and coverage of AraCyc have been extensively enhanced by manual curation. A total of six publications resulted from this work (marked with an asterisk in References).

Prior work on MetaCyc: MetaCyc is a literature-based, multi-species metabolic reference database. It describes experimentally-determined pathways, reactions and enzymes from a multitude of organisms. The MetaCyc data is stored within a frame knowledge representation system called Ocelot[33], which uses an object-oriented data schema. Different types of objects, such as pathways and reactions, are organized into classes. Individual objects, such as a pathway, are represented as frames containing slots that describe attributes of the biological object or relationship among that object and other objects. MetaCyc resides in the software environment Pathway Tools, written in the LISP language. Pathway Tools has three components: 1) the Pathway/Genome Editor, which supports curation, 2) the Pathway/Genome Navigator, which supports querying, browsing, visualization, data analysis, and web publishing, and 3) PathoLogic, which predicts and creates species-specific PGDBs from annotated genomes using a reference database such as MetaCyc. All the PGDBs created by Pathway Tools share the same data schema as MetaCyc, and they can be curated, published, and shared in the same software environment. In 2002, TAIR joined the MetaCyc curation effort specifically to enhance its plant pathway and enzyme coverage. In summary, the number of plant pathways in MetaCyc has grown from 18 (4% of total pathways in release 6.0 in 2002) to 266 (35% of total pathways in release 10.1 in 2006) as a result of our curation effort. This increase includes 167 pathways that occur in Arabidopsis and an additional 81 non-Arabidopsis plant pathways, most of which are involved in secondary metabolism. In addition to curating pathways and enzymes, we significantly improved the way in which plant secondary metabolism and subcellular compartments are organized in MetaCyc. Currently, 114 plant secondary metabolism pathways are organized under eight main subclasses, which are composed, in turn, of 26 subclasses covering all major secondary metabolites reported in plants. Additionally, the cell component ontology has been substantially overhauled. The previous flat list of 35 terms describing subcellular components (mostly of prokaryotic and simple eukaryotic cells) has been structured into an ontology and expanded to include 160 terms[34]. The cell component terms are used in annotating locations of enzymes and transporters.

Prior work on AraCyc: AraCyc (version 1.0) was originally built in 2001[19] using the PathoLogic component of the Pathway Tools and MetaCyc as the reference database. In 2004, AraCyc was rebuilt to incorporate the updated annotations from the Arabidopsis genome resulting in version 2.0[34]. In the 2.0 build, we also made use of the GO annotations that had become available for the Arabidopsis genome[35]. All 7,900 loci annotated to the GO term 'catalytic activity' (GO:0003824) or to its child terms were examined manually. Excluding loci involved in macromolecule metabolism (e.g. peptidases), and loci whose catalytic activity is not involved in metabolism (e.g. transposases), 4,896 loci were included in the input file for building AraCyc 2.0, which resulted in the prediction of 219 pathways. We manually reviewed all of the pathways by consulting the primary literature, review articles, textbooks and experts in specific areas of metabolism. This review resulted in the removal of 82 pathways for which no literature evidence could be found confirming their existence in Arabidopsis. An additional 27 pathways that were not confirmed were nonetheless kept in AraCyc because reactions or metabolites that are unique in these pathways are known to exist in Arabidopsis. In addition, 91 pathways were added and diagrams for 35 pathways were updated. The latest version of AraCyc (2.6) contains 228 pathways of which 201 have literature support. In addition, 1757 enzymes and 1808 genes are annotated to pathways, which is significantly greater than the number of enzymes and genes in MetaCyc (866 plant enzymes and 684 genes). A total of 1572 citations are included in the current version of AraCyc.

Of the 1,401 total reactions in AraCyc, 407 are not yet associated with any enzymes. The pathway hole filling function of the Pathway Tools software was only able to add putative enzymes to 10% of the reactions lacking enzyme information. To identify candidate enzymes for more of these reactions, we applied a series of methods. First, we assembled the attributes of these reactions, including Enzyme Commission (EC) number, reactants, products and related pathway information from AraCyc and MetaCyc. From this information, a set of 228 EC numbers was generated for the purpose of searching the sequence databases UniProt, GenBank, and BRENDA, to retrieve sequences of enzymes known to catalyze these reactions in other organisms. This search yielded 38,586 sequences for 203 reactions with EC numbers, of which 186 reactions had at least five sequences identified in other organisms. Sequences associated to the 186 reactions were used in running BLAST and Hidden Markov Model (HMM) alignments[36-38] to search for homologous Arabidopsis sequences. This search has identified 601 Arabidopsis candidate genes for 124 reactions. Further review of the strength of matching on the basis of expectation values returned by BLAST and HMMer, along with consideration of current gene ontology annotations is being carried out before assigning them to the AraCyc reactions with a computational evidence flag.

The process of literature curation involves a variety of tasks, including: 1) the extraction of accurate pathway diagrams from the literature, 2) the assignment of EC numbers to reactions, 3) the incorporation of metabolite structures, 4) the annotation of enzymes and genes to the reactions, 5) the extraction of enzymatic properties and enzyme regulatory information, and 6) the writing of comprehensive summaries for pathways and enzymes. Typically, a pathway commentary contains a general description of the pathway, its significance, information on its metabolites, as well as regulatory information and rate-limiting steps. Importantly, each pathway and enzymatic reaction is assigned an evidence code (experimental or computational) which, along with literature citations, allows users to rapidly assess the reliability of the information provided. To help curators streamline the literature search process, we have set up the MetSearch literature database, adapted from PubSearch[39]. MetSearch downloads, stores and organizes full text articles describing plant biochemical pathways, enzymes or compounds.

Database availability and usage: AraCyc and MetaCyc data are accessible to the public in a number of ways: 1) searching, browsing, visualization and data analysis through the web interface; 2) complete database download, in BioCyc format and BioPax format under a license agreement (an open database license for AraCyc); 3) compounds dumps in MOL file format and reactions dumps in SBML file; 4) bulk data sets of AraCyc pathways and compounds as tab-delimited text files downloadable through FTP without a license; and 5) third-party databases which incorporate and display the AraCyc data, including MetNet and MapMan. Additionally, we developed PerlCyc and JavaCyc adaptors to interface Perl or Java scripts with the Pathway Tools LISP environment. These adaptors facilitate computational updating of PGDBs using other common computer languages.

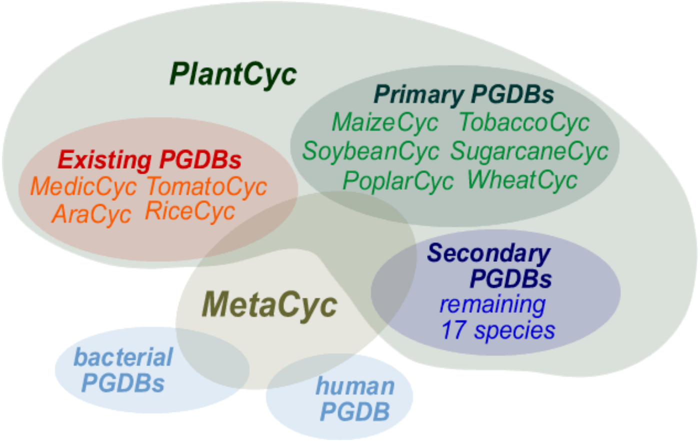
The web site usage of both AraCyc and MetaCyc has increased continuously since the inception of AraCyc, and the beginning of plant pathway curation to MetaCyc. In May 2006, we provided 84,835 AraCyc page views and in the same period the MetaCyc central site provided 33,012 MetaCyc page views. Users may also install AraCyc or MetaCyc on their local computer. Fifty-eight groups have licensed the AraCyc database, and 1700 groups have licensed the MetaCyc database.

Outreach and broader impacts: We have been actively engaged in training scientists to use the plant biochemical pathway resources within AraCyc and MetaCyc through workshops at annual meetings of the American Society of Plant Biology, the international Arabidopsis conference, and the Plant and Animal Genome meeting. We have also presented talks at Plant Metabolomics meetings and a Phytochemical Society of North America meeting, and published our work in peer-reviewed journals[14, 16, 19, 34, 40, 41]. In addition, to leverage the plant community's expertise and enhance curation quality and quantity, in 2006 we held two annotation jamborees at TAIR. In each session, three experts in specific areas of metabolism were invited to interactively work with curators to review curated plant pathways and identify missing information. As a result, 40 missing pathways and 25 out-of-date pathways were identified. We have since established long-term collaborations with these experts (see letters from Drs Facchini, Page, Cahoon, Bar-Peled and Siegler). We are also in close contact with other plant metabolism databases. For example, we provided AraCyc data for incorporation into MetNet[42] and helped Gramene in creating RiceCyc[20]. In addition, we have collaborated with GO on generating the metacyc2go mapping file, which users can use to relate GO terms with MetaCyc pathways and reactions. We have also been in contact with EC in submitting enzymes with missing EC numbers and with BRENDA to create reciprocal hyperlinks.

Context of the current proposal based on prior work: AraCyc is the first example of how the creation of a plant PGDB was jump-started by computational prediction from an annotated genome and quickly increased in quality by manual curation. The pressing need for a high quality biochemical pathway database within the plant research community is evidenced by the growing number of AraCyc users (84,835 web page requests in May 2006). As more genome and EST sequences become available from crop species, there is an ever-increasing interest in annotating those sequences within the context of metabolism by creating species-specific PGDBs for them. Furthermore, the plant community has long been waiting for a central reference metabolic pathway database for all plants to facilitate cross-species analysis and facilitating gene discovery. The proposed work will capitalize on our accomplishments and expertise in creating and curating high-quality databases. Our readiness to tackle the next step in creating the additional plant PGDBs and PlantCyc is based on several factors. The genome function annotation expertise developed along the TAIR project will enable us to efficiently annotate the starting sequences for the creation of a PGDB. Our experience in creating and publishing AraCyc using Pathway Tools software can be readily amplified to the new PGDBs. Our experience in data exchange between AraCyc and other databases will ensure efficient data exchange between PlantCyc and all other PGDBs, and eliminate redundant curation work within individual databases. We will continue to enhance the quality of plant pathway information by extracting literature information using our current literature curation infrastructure. We will also continue our established collaborations with plant metabolism experts and the curators of the 3 other existing plant PDGBs to ensure the highest data quality with minimal duplication of effort.

4. Experimental design

We propose to create a network of plant metabolism databases. At the center of the network will be PlantCyc, which will contain valid plant pathways from multiple plants, supported by experimental evidence for pathways, reactions or enzymes. PlantCyc will be used as a reference database to create plant PGDBs for organisms with substantial sequence data. As each PGDB is built, all of the pathways and enzymes in the new PGDB will be validated and added to PlantCyc, and subsequently curated. Therefore, with each round of PGDB prediction, the quantity and quality of PlantCyc data will be increased. We will leverage the curation teams in other databases interested in different species as well as biochemistry experts interested in specific domains of metabolism. We will build our first set of new PGDBs for plants having the most abundant literature in the area of metabolism. This way, the PGDBs for plants that are less well-studied will benefit from the information already gathered from more intensely studied plants.



Specific Aim 1: Creation of PlantCyc

Initially, PlantCyc will be created by combining AraCyc (created by our group) with the curated information from the three already existing plant PGDBs generated by other groups. Thereafter, PlantCyc will be used to generate new plant PGDBs, which after validation and literature curation will be incorporated back into PlantCyc, thereby iteratively increasing its data content and utility in generating additional PGDBs.

Goal 1: Initiation of PlantCyc: PlantCyc will initially be composed of the entire content of AraCyc along with curated data from other existing plant PGDBs. AraCyc is to date the most comprehensive and accurate plant PGDB available. In addition to AraCyc, three plant PGDBs have already been generated: RiceCyc (*Oryza sativa*), TomatoCyc (*Lycopersicon esculentum*) and MedicCyc (*Medicago truncULata*). Also, a manually curated database exists for soybean, Soybase. To prevent the propagation into PlantCyc of non-plant pathways that were erroneously predicted, we will only incorporate information (pathways, enzymes) from these PGDBs that has been manually curated or adequately validated by those groups as genuine plant pathways. We will collaborate with the groups that have generated these databases in importing the curated pathways into PlantCyc (see letters of support from Drs Mueller, Urbanczyk-Wochniak, Jaiswal and Shoemaker).

Goal 2: Maintenance of PlantCyc: PlantCyc will be updated and maintained in three ways. First, we will exchange data with existing databases that contain curated enzyme and pathway information (see below). Second, we will generate PGDBs for 23 plant species with a significant number of sequences and information about metabolism, and incorporate validated information from these new PGDBs into PlantCyc (Specific Aim 2). Third, we will curate the information in PlantCyc directly using the literature and metabolism experts to update the pathways and incorporate missing pathways (Specific Aim 3). PlantCyc will be released quarterly.

Data exchange with established databases

MetaCyc: MetaCyc data content is regularly updated from a variety of sources such as NC-IUBMB (EC) and chemical databases. We will propagate those improvements into PlantCyc reactions and compounds. In addition, new pathways will be reviewed at each MetaCyc release and imported if relevant to PlantCyc. We will in turn enhance MetaCyc's content by exporting

PlantCyc's experimentally verified pathways and enzymes into MetaCyc (see letter from Peter Karp).

Externally-maintained PGDBs: RiceCyc, MedicCyc, TomatoCyc and Soybase will be maintained externally by our collaborators (see letters from Drs. Mueller, Urbanczyk-Wochniak, Jaiswal and Shoemaker). We will set up information exchange pipelines between these external PGDBs and PlantCyc. They will include new/updated curated genes and proteins, and new/updated pathways. In addition, PlantCyc to PGDBs pipelines will include the propagation of compound and reaction updates (e.g. new and modified EC definitions), and new pathways of potential interest (computational or curated from other species) to be considered for incorporation into the PGDB.

Compound databases: PubChem[43] and KNApSAcK[44] are two important compound repository-databases with slightly different foci. PubChem is a widely used, universal chemical substance database. KNApSAcK on the other hand focuses on natural compounds which are taxonomically classified and referenced with peer-reviewed literature. We will provide PubChem and KNApSAcK links to all of our compounds. PlantCyc's compounds not found in PubChem will be submitted for addition. Finally, KNApSAcK's compounds will be imported into PGDBs in a species-specific manner.

GO: we will contribute to the improvement of GO contents by submitting new biological process (pathways) and molecular functions (enzyme activities) terms to GO ([45]; see letter from Judy Blake). We have also been collaborating with GO to create the MetaCyc2GO mapping file[46] which maps reaction, and pathway IDs in MetaCyc to GO function and process terms, respectively. We will now focus on the creation of a similar mapping file: PlantCyc2GO.

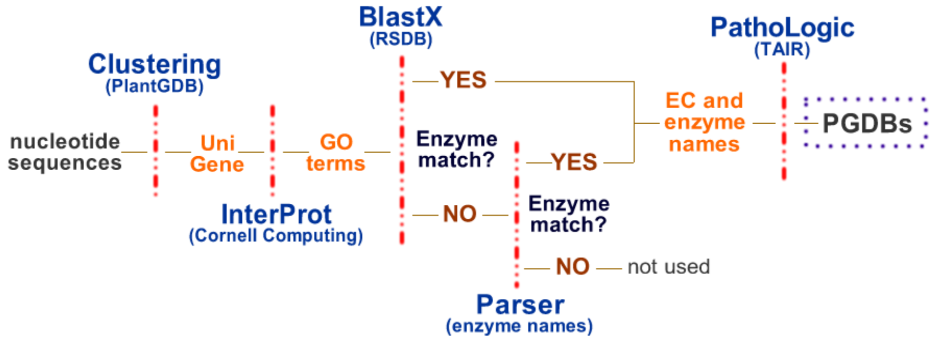
Specific Aim 2: Creation of plant species-specific PGDBs

We will focus on generating new PGDBs for 23 organisms. Our choices were based on a number of criteria including the size of the gene/EST sets, the utilitarian and agronomical value of the individual species, as well as taxonomic interest. High value was given to species with potential impact on bioenergy development (e.g. canola, switchgrass, sunflower, sugarcane), renewable fiber (e.g. cotton, pine, poplar), animal and human food (e.g. sweet orange, wheat, wine grape, sorghum, trefoil, lettuce, apple, banana) and pharmaceuticals for human health (e.g. soybean, cocoa, tobacco). The organisms that we chose can be divided into two categories: 1) those whose genomes are sequenced or underway such as *Populus trichocarpa* (poplar), *Physcomitrella patens* (a moss), *Lotus corniculatus* var japonicus (Japanese bird's foot trefoil), *Selaginella moellendorfii* (a fern), and the salt tolerant *Thellungiella halophila* (salt cress); and 2) organisms with a significant number of EST/cDNA sets such as *Brassica napus* (rape)*, Capsicum annuum (chili pepper), Citrus sinensis* (sweet orange)*, Coffea sp. (coffee), Glycine max* (soybean)*, Gossypium* sp. (cotton), *Helianthus annuus* (sunflower), *Malus X domestica* (apple)*, Nicotiana tabacum* (tobacco), *Panicum virgatum* (switchgrass), *Pinus taeda* (loblolly pine), *Saccharum officinarum* (sugarcane), Solanum tuberosum (potato), *Sorghum bicolor* (sorghum)*, Theobroma cacao* (cocoa), *Triticum aestivum* (wheat), *Vitis vinifera* (grape), and *Zea mays* (maize).

We intend to share these PGDBs with expert plant communities best suited to curate and maintain them. Several groups have already expressed interest in 'adopting' some of the PGDBs that we will be generating. These include PGDBs for grasses such as wheat and sorghum (Gramene, see letter from Dr. Jaiswal); chilli pepper, potato and coffee (SGN, see letter from Dr. Mueller); grape (Grape Consortium, see letter from Dr. Cushman); maize (MaizeGDB, see letter from Dr. Lawrence); and soybean (see letter from Dr. Shoemaker). The PGDBs we create will be made available to those groups for expert curation. We will additionally provide frequent updates for those PGDBs to reflect the continual discovery of new genes and UniGenes, as well as improvements of sequence annotations (see below, Goal 4). We will coordinate curation efforts between these groups but will allow them to take charge of curation for the species they have chosen to adopt. In return, our collaborators will provide PlantCyc with manually curated information on enzymes and pathways for those species. This collaborative framework will avoid redundant curation and homogenize annotation and curation practices between groups interested in plant metabolic processes.

Goal 1: Annotation of the sequences

Step 1. Retrieval of gene sequences: PGDBs will be generated using either gene sequences retrieved from genomic projects or from mRNA sequences. mRNA sequences will first be clustered into UniGenes. A number of groups (PlantGDB, NCBI-UniGenes, TIGR, GénoPlante) already provide the community with reliable UniGene clusters. PlantGDB focuses on the clustering of plant EST and cDNA sequences[47] and frequently updates its clusters to reflect the most current information available from GenBank. Since PlantGDB already provides quality, quantity and frequent updates in its clustering work, we will collaborate with this group and use their clustered UniGenes as starting material for the creation of PGDBs (see letter from Dr. Brendel). For the genome PGDBs, gene sequences will be retrieved from GenBank.



Step 2. Identification of catalytic domains: InterProScan[48] detects the presence of known functional domains for a given protein sequence. These domains (or their absence) help to predict a protein's catalytic activity. Due to the large volume of data to be analyzed, we will run InterProScan (IPRSCAN) through the Cornell Computing Facilities (CBSU) (see letter from Qi Sun). IPRSCAN accepts nucleotide sequences as inputs, and processes the sequences with six-frame translations. The output of IPRSCAN contains a detailed description of the domain found in the sequences including InterPro domain IDs and associated GO function term. All sequences (genes and UniGenes) will be processed using IPRSCAN. The gene sequences will be annotated with the resulting e-value and status of the domain prediction.

Step 3. Functional annotation of genes: The creation of PGDBs by PathoLogic requires the input sequences to be functionally annotated as enzymes. Many of publicly available sequences have already been functionally annotated (e.g. TIGR Gene Index, GénoPlante), yet the annotation protocols used by the different groups are heterogeneous and of variable quality. In order to obtain consistent annotations of reliably high quality, we will annotate all gene sequences using our in-house functional annotation pipeline.

Step 3a. BlastX against Reaction Sequence Database (RSDB): The first round of annotation will match sequences against enzymes with known, well-characterized catalytic activities. This will be achieved by running a BlastX search[49] of our genes and UniGenes against an Enzyme database, the Reaction Sequence DataBase (RSDB), which our group has started compiling. RSDB catalogues protein sequences gathered from many enzyme sources including UniProt, PDB, and MetaCyc. Each enzyme sequence in RSDB is associated with 1) the reaction(s) it catalyzes (using MetaCyc IDs), 2) if available, the EC numbers of these reaction(s), 3) the enzyme's catalytic activity(ies), 4) the catalytic domain(s), and 5) the associated GO annotations. We will blast the gene sequences against the RSDB sequences. The query sequences that match at an e-value of 1 x e^{-10} or less with an overall sequence match of 80% of the query sequence will be retrieved. The predicted catalytic activity of sequences matching the RSDB entries will be further analyzed to confirm the presence of expected catalytic domains as defined in RSDB for the entry to which the sequence matches. For each matched sequence, a description of the Blast e-value will be provided, along with the InterPro domains (including missing domains compared to the matched RSDB entry), enzyme name and reaction association (including EC numbers when available) of their matched entries. This information will be valuable to users wanting to assess the reliability of the predictions. Sequences with matches better than the defined cutoff values will be annotated to the matching RSDB enzyme names and EC numbers.

Step 3b. EC numbers and Enzyme name parsing from GO function terms: If sequences fail to produce significant matches with RSDB at the previous step, we will attempt to further identify their putative catalytic activities. Indeed, EC numbers have a corresponding function GO term. The associations between these two nomenclatures are being regularly compiled in a mapping file called EC2GO[50]. This file and the GO function terms obtained from the IPRSCAN analysis (see Step 2) will be used to try to retrieve EC numbers for each remaining sequence having failed Step 3a. Only Interpro domain predictions with an e-value superior to 1E-10 and a 'True' status (IPRSCAN's suggested quality cutoff) will be assessed. Sequences with a positive match will be assigned the corresponding EC number and enzyme activity, as well as the details of the computational prediction (method, e-value and status). If no EC numbers can be matched by this method, the IPRSCAN-generated function GO terms (if any) will be parsed to try to identify an enzyme name. Indeed, GO terms sometimes describe enzymes that have not yet been assigned an EC number; those names can be used to map to reactions using PathoLogic (see below). All enzymes with a GO function term that is a child of 'catalytic activity' will remain in the database. Other entries that have failed through all the steps will not be used at this time.

Goal 2. Creation of new PGDBs

PGDBs will be created in the model of AraCyc by running the PathoLogic component of Pathway Tools software. The parsed enzyme names and EC numbers for sequences annotated in the above steps will serve as inputs for PathoLogic. PlantCyc will be used as the first reference database. PlantCyc will only contain pathways that have been validated for existence in plants and will therefore generate a plant PGDB. We will prioritize the organisms for generating a PGDB based on the amount of gene sequence data and metabolic information available from the literature. The species of highest interest using these two criteria are maize, poplar, sugarcane, soybean, tobacco and wheat. These will be the first ones to be used for PGDB creation and incorporation into PlantCyc. Although PlantCyc will be the primary reference database used in the generation of a plant PGDB, a few additional pathways might be identified using MetaCyc. Therefore, MetaCyc will be used as a secondary reference database. If any additional pathways are identified from the MetaCyc-based PGDB build, the relevant pathways will be brought into the new PGDB only upon validation.

Upon creating a PGDB for these species, the predicted pathways will be validated using the pathway validation criteria defined during the AraCyc build[34]. According to these criteria, a pathway is kept in a PGDB when: 1) it is referenced in the literature for the studied species; or 2) it has enzymes predicted to catalyze reaction(s) unique to the pathway schema; or 3) metabolites unique to the pathway exist in the species studied. Pathways failing to meet one of these criteria are deleted as false positives. Following validation, the PGDB will be considered publishable. A website will be set up for that PGDB and will be release to the public.

Goal 3: Incorporation of new PGDBs into PlantCyc: Newly predicted and validated pathways will be incorporated into PlantCyc. All pathways of a PGDB missing from PlantCyc will be added (those predicted from MetaCyc and validated), improving the comprehensiveness of its PGDB predictability. All proteins and genes from the PGDB will be transferred to PlantCyc using the bulk pathway export/import tool provided with the Pathway Tools software. This function contains a variety of integrity and consistency checks assuring that all objects (compounds, reactions, pathways) are being merged properly. Once incorporated into PlantCyc, the new pathways will be curated from the literature (Specific Aim 3).

Goal 4. Updating PGDBs: Continual expansion of mRNA sequence collections and updated annotations of genomes require the PGDBs to be updated. The first tier PGDBs (maize, poplar, sugarcane, soybean, tobacco and wheat) will be updated biannually while other PGDBs will be updated approximately once a year. This timeline is flexible as we intend to accommodate a potential influx of new information from public databases. For example, PlantGDB currently updates all its clusters quarterly, following GenBank releases[51]. PGDBs will be updated ahead of schedule when the number of UniGenes in PlantGDB has increased by at least 20% for PGDBs built with less than 50,000 sequences, 10% for PGDBs built with 50,000 to 80,000 sequences, and 5% for PGDBs built with more than 80,000 sequences. Priority will be given to organisms with the largest changes in enzyme annotation. In addition, we will also update RSDB on a quarterly basis (see Goal 1- Step 3a). PathoLogic will be run as described in the previous section and the updated PGDBs will replace the previous version which will be archived. Additions, deletions and modifications in UniGene assignments to reactions will be propagated to PlantCyc. Following this procedure, all manually curated information for a given species found in PlantCyc will be transferred into the respective PGDB. In order to remain consistent, UniGene IDs will remain identical between versions. In collaboration with PlantGDB, changes in UniGenes assignment to reactions will be disclosed for easy tracking (see letter from Dr. Brendel).

Specific Aim 3: Improve PlantCyc and the PGDBs

We will improve the content of PlantCyc by updating the existing pathways with an emphasis on in-depth enzyme curation and adding pathway variants and missing pathways from the literature. We will follow the high quality standards of curation that have been developed for AraCyc and MetaCyc. The information used for the curation of the pathways will be gathered from a variety of sources. First, new pathways will be identified through literature mining: peer-reviewed articles will be retrieved from various resources, such as PubMed, Scirus[52] and our own database MetSearch (see Preliminary results). Second, resources identifying compounds will also be exploited to identify potential targets for pathways missing from the database. These include the analysis of large-scale metabolic profiling datasets[53-55], consulting species-specific compound databases such as KNApSAcK[44] as well as the study of metabolite-oriented journals such as Phytochemistry, Journal of Natural Products and Natural Product Reports. Finally, we will continue to seek the contributions of experts who will review and add new pathways in their field of specialty. Involvement of the scientific community will accelerate the improvement of PlantCyc's content and ensures the best quality control through such a community-based review system (more details in Specific Aim 4).

Goal 1. Updating existing pathways

Enhance pathway comments: Primary metabolism is central to a large number of bioengineering projects. For instance, pathways classified under carbohydrate metabolism extensively contribute to biomass production which in turn can be converted to sustainable biomaterials such as biofuels[56, 57] . Although plant primary metabolic pathways are well represented in AraCyc (and therefore PlantCyc), in-depth curation of its enzymes, genes and pathway comments remains largely incomplete. Out of the 228 pathways represented in AraCyc, 95 (41%) primary metabolism pathways need to be updated with pathway comments.
Enhance pathway variant curation: In the course of evolution, plants developed different routes for metabolizing compounds. These pathways variants can be identified through in-depth

curation of corresponding enzymes and associated reactions in different species. For example, cholin, a fundamental metabolite of plant membrane phospholipids, is synthesized in plants via three different routes. The conversion of the common precursor ethanolamine to choline occurs either on the level of free bases such as in castor bean[58], phospho-bases as shown for spinach[59], or phosphatidyl-bases which has been demonstrated in soybean[60]. We will curate such variant routes for pathways. These variants will be used to create additional super-pathways. A super-pathway is defined as an aggregation of two or more pathways that are related in some way and is useful to present the overall biosynthesis or degradation of compounds of interest.

Enhance protein curation: Previous curation of protein information has been limited, particularly for enzymes. For each pathway, only enzymes from one or two representative species were curated. Here we plan to curate all known plant enzymes for pathways of importance to biofuel production and human health. We will also start curating metabolite transporters, a domain that has not been tackled in the previous plant pathway curation effort.

Enzyme functional properties: Knowing an enzyme's catalytic properties such as substrates/alternative substrates, Km and optimum pH values, cofactors, activators, and inhibitors is critical to many aspects of research and applications, from gene discovery to metabolic engineering. We plan to curate all known plant enzymes for pathways which are of potential interest for biofuel production, food nutrition improvements, or medicinal applications. Curation of individual enzymes for functional properties should be straightforward. First, publications containing this type of information can easily be identified and flagged as highly relevant. To this end, we will use MetSearch to identify all plant articles potentially containing enzyme property information, for example articles containing the keyword 'purification' or 'characterization' in the title. Secondly, functional properties are usually described in specific sections of an article and can be easily identified and extracted. In addition to our own efforts in semi-automated extraction of enzyme information, we will also consult existing protein databases such as Brenda[61] which already contains much enzyme property information.

Metabolite transporters: We plan to expand protein curation beyond enzymes to include transporters. Plant biochemical pathways often span multiple subcellular locations or even different cell types which requires shuffling of intermediate metabolites between compartments. A pathway will not be considered complete without curation of all the transporters involved. The Arabidopsis genome alone contains 1250 loci annotated as transporters, of which 210 are based on experimental data from the literature. We will begin transporter curation with these 210 Arabidopsis transporters, after which transporters from other plants will be curated. The curation of transporters will have two main foci: the identity of the transported metabolites and the compartments to/from which translocation occurs. In the Pathway Tools environment, transporters can readily be curated and displayed in a way similar to enzymes.

Goal 2. Add new pathways: The continued addition of new pathways to PlantCyc is central to increasing its breadth. The highest priority will be given to plant pathways that have significant impact on human health and bio-energy production[62, 63]. In addition, we will curate well characterized enzymes catalyzing reactions that are not part of a pathway. Our experience in pathway curation indicates that a focused investigation of orphan reactions also helps identify missing pathways. Pathways to be added can largely be grouped into three categories, primary metabolism, secondary metabolism and hypothetical pathways.

Primary and secondary metabolism: Under primary metabolism, we will curate pathways for carbohydrates, fatty acids, lipids, and vitamins. Examples of pathways that we plan to add include synthesis and degradation pathways for molybdenum cofactor, cardiolipin, unusual fatty acids, rhamnogalacturonana I and II, galacturonan and xyloglucan. In addition, it has been demonstrated in *Pinus taeda* that the biosynthesis of cellulose is coregulated with lignin formation by repressing 4-coumaroyl ligase[64] paving the way for metabolic engineering of cellulose that can be processed into transport fuels[65]. Switchgrass[66] and rice[67] also possess a high capacity for producing lignocellulosic material such as lignin, cellulose and hemicellulose, which has been successfully converted into ethanol. The curation of the underlying pathways is of high relevance to metabolic engineers.

In contrast to primary plant metabolism, biosynthesis and degradation of secondary (specialized) metabolites are less well-represented in plant metabolic databases. The number of identified secondary products in plants is constantly growing[68]. In Arabidopsis alone, a plant that is known to produce relatively few secondary metabolites, the number has increased from 36 pathways organized into four major compound classes[69] to 170 pathways in seven classes[70]. Many end products of these pathways are of considerable interest for researchers and metabolic engineers because of their substantial impact on human health, agriculture and plant specific functions such as defense, scents and pigmentation[71]. Therefore, we will concentrate on plant secondary metabolite pathways that encompass those areas and are reported for the species we intend to include in PlantCyc. For example, soybean (*Glycine max*) contains plant natural products such as pterocarpans (isoflavonoids) that possess potential health benefits[72, 73]; terpenoid alcohols isolated from *Helianthus annuus* have been shown to inhibit tumor formation[74]; and flavan-3-ols isolated from *Theobroma cacao* act as antioxidants and have attracted interest regarding cardiovascular health[75].

Hypothetical pathways: Currently, MetaCyc only contains pathways that have been experimentally verified, largely through the isolation of enzymatic activities. However, a number of plant pathways, such as those involved in the metabolism of complex secondary metabolites, lack enzyme and gene information and are therefore absent from MetaCyc. For instance, in the biosynthesis of sterols, many steps are predicted from semi-synthetic preparations of putative intermediates. Similarly, the multi-step biosynthesis of the potent anti-tumor drug taxol (found in Pacific yew, *Taxus brevifolia*), in spite of having been achieved synthetically, is still very poorly understood[76, 77]. In such cases, pathway schemata contain many hypothetical semi-synthetic steps mixing paper chemistry, organic syntheses and radioactive tracer labelling, but seldom enzymatic activities. In PlantCyc we will accommodate the incorporation of those hypothetical pathways with appropriate documentation of evidence, as was frequently requested by users of AraCyc and MetaCyc. These additions will enhance the usefulness of PlantCyc as an educational and research tool.

Data Access: PlantCyc and individual PGDBs created and curated in this proposed work will be made freely accessible to the public following the same methods used for AraCyc: 1) web interface, 2) complete database downloads under a simple open Database license, 3) bulk data sets as downloadable text files, and 4) third-party databases which incorporate and display PlantCyc data. From the PlantCyc and individual PGDB web sites, users will be able to easily access all data through simple search and browse options, conduct cross-species comparison of the pathways, or analyze their own large scale experimental data by overlaying them on the metabolic network using the Pathway Tools' build-in OmicsViewer functionality. Complete database downloads will allow for the installation of a local copy of PlantCyc, or any species-specific PGDB. These can be used for a variety of purposes including computation or integration of users' own data. As for AraCyc, users wanting to license complete database copies will sign an online license form and, in return, be emailed detailed instructions on FTP access and database installation procedures. Finally, we will also offer bulk datasets for pathways (tab-delimited text files), compounds (MOL file, and tab-delimited text file), and reactions (SBML file) through license-free FTP download. Other public databases including MetNet and MapMan have incorporated AraCyc pathways and made the data available on their web sites. Similarly, any new data curated in PlantCyc will be made available on these web sites (see letters of collaboration from Drs. Wurtele and Stitt). In addition to the bulk data files in text and XML, we will generate files for visualizing the pathways of a PGDB as a network. Current views of individual pathways are not convenient for analyzing the effects of genetic or biochemical perturbation to the entire metabolome of a cell. In order to visualize and analyze the whole metabolic network, we will combine reactions from all pathways in a PGDB and model them as a single directed graph. We will partition the resulting network with a clustering algorithm based on the traffic of connections between nodes. In this method, edges with the highest betweenness centrality are repeatedly removed until all components are of a specified size. We will generate the resulting network data into formats that can be read by the open-source interaction visualization and analysis software, Cytoscape[78].

Our anticipated release schedule for PlantCyc and the PGDBs is as follows: Year 1) release of PlantCyc; Year 2) sequential release of the 3 first-tier PGDBs (maize, poplar and soybean); Year 3) release of the remaining 3 first-tier PGDBs (wheat, tobacco and sugarcane) and 3 secondary PGDBs; Year 4) release of 7 secondary PGDBs and 5) release of final 7 PGDBs. After its initial release in year 1, PlantCyc and AraCyc will be released quarterly, primary PGDBs biannually, and secondary PGDBs annually.

Specific Aim 4: Outreach and Broader Impacts--Create an alliance of plant metabolism communities

Developing a comprehensive infrastructure for a large domain of knowledge is difficult to achieve by one group. We aim to foster cooperation, sharing of resources and expertise ranging from genome annotation databases, metabolism databases, and individual scientists in biochemistry by creating a set of plant metabolic databases and engaging experts in specific domains to participate in their improvement. The resulting network of databases will not only facilitate research, but also provide an up-to-date, easily accessible set of teaching and educational resources for students.

Goal 1. Create a network of community annotation: We intend to initiate a broad network within the plant metabolism community to ensure that data in PlantCyc and PGDBs meets high standards to benefit the goals of all plant biologists and students involved in research on plant metabolism. We will use Pathway Tools' newly implemented author crediting system to acknowledge the contribution of users to PlantCyc. Collaboration with other databases who will 'adapt' a PGDB will occur through established data exchange pipelines as mentioned in specific aim 2. Other types of shared annotations are described here.

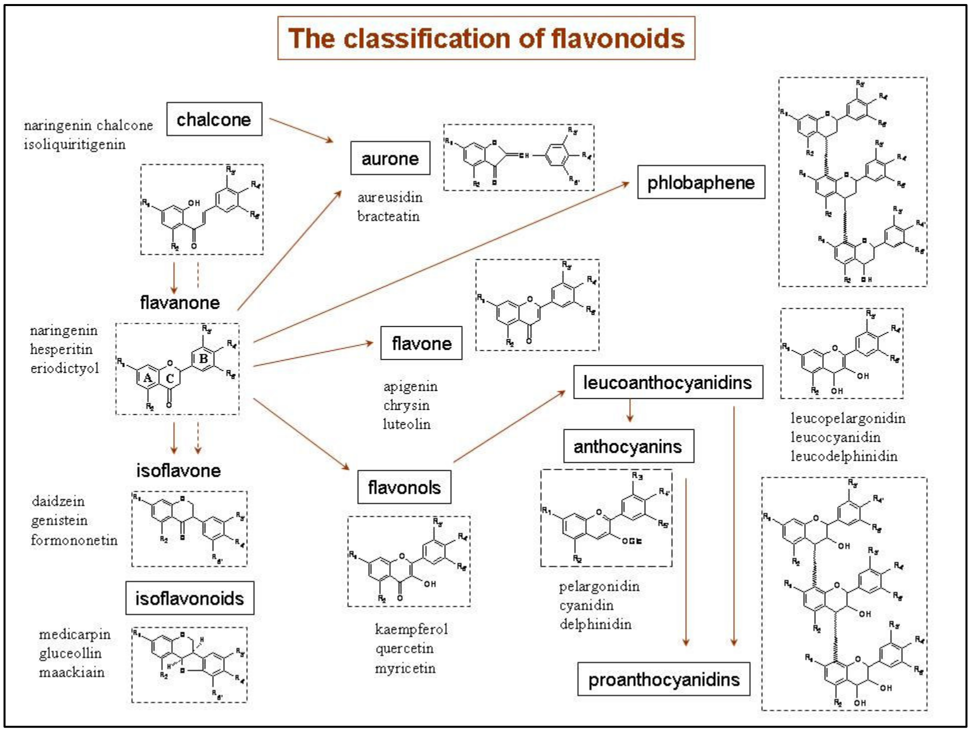
Plant metabolism jamborees: A promising approach is the organization of plant metabolism oriented jamborees. Two pilot jamborees held at Carnegie (see Preliminary results) not only enhanced the content of both AraCyc and MetaCyc with high-quality data, but also increased the awareness of those databases to the community. We will continue to organize such jamborees on a regular basis, on a variety of topics spanning both primary and secondary metabolism. We plan on organizing 2 jamborees a year. The first two jamborees will focus on primary metabolism on domains such as cofactors, amino acids, nucleotides and nucleosides and plant hormones.

Submission of pathways and metabolic data: Currently, people can submit data through two media: 1) through an online submission Excel form[62] and 2) via a feedback page through which users can be easily make suggestions to our curators. We will experiment ways to encourage user submission. For example, enhance the Excel form to have all PlantCyc reaction and enzyme information pre-filled for each existing pathway so that it is easier to submit corrections or missing data.

Outreach: The overall goal of our outreach plan is to spread knowledge about PlantCyc. We will continue to encourage the scientific community to participate in the effort of making a high-quality and widely used plant metabolic database. We plan to visit relevant national and international plant metabolism conferences such as PSNA annual meetings, ASPB (Plant

Biology), annual meetings of the Groupe Polyphenol, International Conference on Plant Metabolomics, and Plant Genomics European Meetings. We will actively participate in those meetings in form of workshops, lectures and posters. The progress over time of PlantCyc and research conducted using the DB will be published in applicable journals to demonstrate the utility of the DB. The outcome of metabolism-focused jamborees will be made public to leverage on the feedback of the interested community.

Goal 2. Create a board of editors: Accuracy, consistency and currency of the released data define the value of the DB for the plant metabolism community. We plan to set up a control system that warrants the quality of the published data on our website by creating a Board of Editors to act as auditors and advisors on the project. The Board will be composed of one or two scientists per area of specialty of which there are currently 14 (six for primary and eight for secondary metabolism). Currently, six outside experts have agreed to work as editors covering both primary and secondary plant metabolism (see letters of support from Drs. Facchini, Page, Cahoon, Bar-Peled and Siegler). In addition four more scientists have expressed their interest to contribute as editor or reviewer of the contents in our DB. Upon funding, we will complete the Board of Editors with additional scientists willing to participate in the quality control of metabolic data presented in PlantCyc. We have compiled a list of potential experts (over 25) that cover the remaining domains in primary and secondary metabolism. The Board members will review pathways for correctness and completeness, and act as a liaison with the scientific community to identify researchers best suited to improve PlantCyc and the plant PGDBs.



Goal 3. Education: PlantCyc's intent is to reach beyond pure research and applications for metabolic engineers. We also consider its value as a reliable and comprehensive educational tool. We will contact and encourage lecturers and teachers to use PlantCyc in their courses. We plan the integration of new features into the DB enhancing the educational value of the DB for professors, students and the interested publicity. For example, we will provide concise overview diagrams of the main plant compound classes. Those tutorials will include chart information as exemplified in the diagram above with the overview of flavonoids. These overviews will present graphical, interactive maps of the more complex compound classes such as flavonoids, terpenoids and alkaloids, where each compound class will be hyperlinked to its corresponding pathways in the database. Also, we will set up a glossary for metabolic terms and concepts to facilitate the use of the DB. In addition, we will develop and integrate tutorials for PlantCyc in a similar format to those currently used for AraCyc.