MedicCyc: a biochemical pathway database for Medicago truncatula

Ewa Urbanczyk-Wochniak and Lloyd W. Sumner*

The Samuel Roberts Noble Foundation, Inc., Ardmore, OK 73401, USA

Received on October 26, 2006; revised on January 29, 2007; accepted on February 1, 2007

Advance Access publication March 7, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: There is an imperative need to integrate functional genomics data to obtain a more comprehensive systems-biology view of the results. We believe that this is best achieved through the visualization of data within the biological context of metabolic pathways. Accordingly, metabolic pathway reconstruction was used to predict the metabolic composition for Medicago truncatula and these pathways were engineered to enable the correlated visualization of integrated functional genomics data.

Results: Metabolic pathway reconstruction was used to generate a pathway database for M. truncatula (MedicCyc), which currently features more than 250 pathways with related genes, enzymes and metabolites. MedicCyc was assembled from more than 225 000 M. truncatula ESTs (MtGI Release 8.0) and available genomic sequences using the Pathway Tools software and the MetaCyc database. The predicted pathways in MedicCyc were verified through comparison with other plant databases such as AraCyc and RiceCyc. The comparison with other plant databases provided crucial information concerning enzymes still missing from the ongoing, but currently incomplete M. truncatula genome sequencing project. MedicCyc was further manually curated to remove non-plant pathways, and Medicago-specific pathways including isoflavonoid, lignin and triterpene saponin biosynthesis were modified or added based upon available literature and in-house expertise. Additional metabolites identified in metabolic profiling experiments were also used for pathway predictions. Once the metabolic reconstruction was completed, MedicCyc was engineered to visualize M. truncatula functional genomics datasets within the biological context of metabolic pathways.

Availability: freely accessible at http://www.noble.org/MedicCyc/

Contact: lwsummer@noble.org

1 INTRODUCTION

A comprehensive and accurate understanding of both primary and secondary metabolism is fundamental to the knowledge of plant biology. Although primary metabolism is one of the best reconstructed networks within biological systems, our understanding of secondary metabolism and its metabolic regulation is still quite limited. Improved functional genomic technologies such as transcriptomics, proteomics and metabolomics offer stimulating opportunities for a more thorough understanding of metabolism and we currently utilize an integrated functional genomics approach to study the model legume Medicago truncatula (Achnine et al., 2005; Broeckling et al., 2005; Lei et al., 2005; Sumner et al., 2003; Suzuki et al., 2005). It is presumed that the model legume M. truncatula will be the fourth plant genome to be fully sequenced (Town, 2006; Young et al., 2005), and as of January 2007, ~190 Mbp of euchromatin sequence has been acquired (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago) representing approximately 25 000 predicted genes. At the same time, the collection of ESTs (http://www.tigr.org/tigr-scripts/igl/T_index.cgi?species=medicago) has increased to 18 238 non-redundant singleton ESTs and 30 000 non-redundant tentative consensus (TC) sequences (MtGI Release 8.0). With the continually mounting sequence of M. truncatula, release of the fully sequenced Arabidopsis (Arabidopsis Genome Initiative, 2000), rice (International Rice Genome Sequencing Project, 2005), and black cottonwood (Tuskan et al., 2006) genomes, and the initiation of many other plant sequencing projects; there is a fast growing need for a better understanding of the functional aspects of sequenced genomes. This demand can be best met through an integrated functional genomics approach, which typically includes a combination of transcriptomics, proteomics and/or metabolomics. Genomic and expressed tag sequencing of M. truncatula (Young et al., 2005) have provided resources for the assimilation of custom and commercial cDNA and oligonucleotide high-density microarray chip sets for transcriptome analysis (Barnett et al., 2004; Covitz et al., 1998; Kuster et al., 2004). Similarly, M. truncatula nucleotide sequencing and advances in mass spectrometry have enabled large-scale profiling of proteins (Gallardo et al., 2003; Imin et al., 2004; Lei et al., 2005; Mathesius et al., 2001; Watson et al., 2003). The development of metabolomics is more recent (Fiehn, 2002; Sumner et al., 2003; Weckwerth, 2003), and the number and variety of metabolites that can be routinely measured have remarkably increased over the last several years (Broeckling et al., 2005; Chen et al., 2003; Huhman and Sumner, 2002; Huhman et al., 2005; Liu et al., 2003).

Biochemical pathway maps composed of genes, proteins and metabolites (Zhang et al., 2005) are powerful tools around which one can compile the biological context of
functional genomics datasets. Currently, several legume databases exist and include The Legume Information System (http://www.comparative-legumes.org/) (Gonzales et al., 2005), The Medicago truncatula consortium Database (MtDB, http://www.medicago.org/MtDB) (Lamblin et al., 2003) and the Medicago Gene Index (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago) (Cannon et al., 2005). Of these, only the TIGR database provides linkages of tentative consensus sequences to metabolic pathways, and to the best of our knowledge, none of the databases listed are currently capable of visualizing functional genomics data (e.g. expression data obtained by microarray analysis) within the context of biological pathways. However, such databases do exist for other species. For example, the visualization tool MapMan has been developed for the display of metabolite and transcript data onto metabolic pathways of Arabidopsis and tomato (Thimm et al., 2004; Urbaczynk-Wochniak et al., 2006; Usadel et al., 2005). Similarly, a commercial software has been modified to visualize Arabidopsis transcript and metabolite data (Lange and Ghassemian, 2005). To the best of our knowledge, there is no dedicated tool or publicly accessible database designed for accommodating or visualizing M. truncatula functional genomics data. As a result, we have developed MedicCyc, a biochemical pathway database for M. truncatula which can be used as a reference for M. truncatula as well as other legume species and is freely accessible at http://www.noble.org/MedicCyc/.

MedicCyc was constructed using MetaCyc, a database composed of all experimentally determined biochemical pathways for small molecule metabolism (Caspi et al., 2006; Keseler et al., 2005; Krieger et al., 2004), and the related Pathway Tools software (Karp et al., 2002; Paley and Karp, 2002). MetaCyc consists of reactions, enzymes, metabolites, feedback regulation information and genes that encode enzymes accumulated from a large number of species to enable the prediction of pathways in new species. The current version of MetaCyc (10.0) contains approximately 700 pathways from more than 600 species, ranging from microbes to plants and human (http://metacyc.org/). The Pathway Tools software allows automated generation of a pathways database through functional assignment of genes and manual editing of pathways using a graphical user interface. Currently, more than 90% of pathways integrated into MetaCyc are manually curated and include literature citation and species information (Caspi et al., 2006). To date only two pathway genome databases (PGDBs) have been created for plants. The first was AraCyc for Arabidopsis thaliana (http://www.arabidopsis.org/tools/aracyc/) (Mueller et al., 2003; Zhang et al., 2005) and more recently, the Gramene Community (a Resource for Comparative Grass Genomics) reconstructed the biochemical pathways for rice and created RiceCyc—a web-based tool for viewing gene annotations mapped to various biochemical pathways in two model plant species, Arabidopsis and rice respectively (for more details visit http://www.granome.org/pathway/). The utilization of a common ‘Cyc’ database format provides a consistent and more equitable platform for the comparison of reconstructed plant pathways between Medicago and other available PlantCyc databases. This is possible through the utilization of a Pathway Tools comparative module. Pathway comparisons between plant species can also reveal current gaps in the knowledge of Medicago metabolism.

MedicCyc, a curated and integrated pathway database for M. truncatula, was developed to enable further study of metabolism and functional genomics in M. truncatula. The goals of this project were: (i) to use pathway reconstruction for predicting the metabolic composition of M. truncatula; (ii) to provide a platform for visualization of the integrated functional genomics datasets. Furthermore, the long-term goals are: (iii) to enable the prediction and annotation of unknown genes through comparison with other sequenced plant genomes and detected metabolites; and (iv) to provide an additional resource for the metabolic engineering of legume metabolism and the enhancements of desired traits.

2 METHODS

2.1. MedicCyc construction using transcript annotation

MedicCyc was constructed using metabolic annotated transcripts selected from The Affymetrix GeneChip® Medicago Genome Array. The Medicago Genome Array was designed specifically to monitor gene expression in Medicago truncatula, Medicago sativa and additionally the symbiotic organism Sinorhizobium meliloti (http://www.affymetrix.com/products/arrays/specific/medicago.affx). Sequence information for this array was selected from diverse data sources including the TIGR M. truncatula gene index (MtGI Release 8.0), automated gene predictions from The International Medicago Genome Annotation Group (IMGAG, http://www.medicago.org/genome/IMGAG.php) (Town, 2006), gene predictions from the S. meliloti genome, and M. sativa EST information released by TIGR. The array consists of more than 32 167 Medicago truncatula EST/mRNA-based and chloroplast gene-based probe sets as well as 18 733 M. truncatula IMGAG and phase 2/3 BAC prediction-based probe sets, which are believed to cover the majority of the transcribed part of the M. truncatula genome. It has been predicted that ~25% of the Arabidopsis genome encodes metabolic enzymes and the value may be similar or even higher for legumes (Bell et al., 2001). Accordingly, 4400 transcripts were pre-selected from Medicago Genome Array based upon the corresponding annotations that suggest that these genes are involved in metabolism.

Due to the current and incomplete stage of M. truncatula genomic sequencing, genes and corresponding transcripts have not been completely associated with chromosomes. Therefore, MedicCyc transcripts were associated as arbitrary ‘groups’ which do not represent the typical chromosomal grouping utilized by Pathologic. After completion of the genomic sequencing, all transcripts used to build MedicCyc will be linked and grouped to specific chromosomes. The analog annotations were formatted into a PathoLogic-specific set according to the documentation for Pathway Tools (Keseler et al., 2005) and used for the first MedicCyc database construction. The annotations were edited manually to remove extraneous words and characters that could interfere with the enzyme name-matching software, which was analogous to the process used for construction of AraCyc (Mueller et al., 2003). Enzymes labeled as ‘putative’ or ‘similar to’ were also included in the dataset. Additionally, the selected transcripts were coupled with corresponding EC (Enzyme Commission) numbers, which were used to map the transcripts onto pathways and to avoid problems with enzyme synonyms. The initial MedicCyc database was generated using the PathoLogic Pathway Prediction module of Pathway Tools ver. 10.0. The initial Medicago pathways (MedicCyc ver. beta) were inferred using MetaCyc as a reference database of metabolic pathways, and later MedicCyc ver. 1.0 was curated using AraCyc and RiceCyc PGDBs as co-reference databases.
2.2 MedicCyc construction using previously detected metabolites

To evaluate the completeness of the _M. truncatula_ metabolic reconstruction, we compared the compounds in MedicCyc with compounds previously detected and identified in metabolic profiling experiments (Broeckling et al., 2005; Chen et al., 2003; Huhman and Sumner, 2002; Huhman et al., 2005; Liu et al., 2003). Forty-two compounds reported in these studies were not found in the initial version of MedicCyc (Table 1). Missing compounds included secondary metabolites recently reported in _M. truncatula_, e.g. different isoflavonoids (Deavours et al., 2006; Farag et al., 2007). The absent metabolites were manually inserted into the database and most of them were used to reconstruct new metabolic pathways, e.g. triterpene saponin biosynthesis. Continuing metabolic profiling experiments will further extend these pathways as our knowledge of secondary metabolite biosynthesis and degradation is still far from complete.

### Table 1. Manually added compounds into MedicCyc (1.0) database

<table>
<thead>
<tr>
<th>Compound</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayogenin</td>
<td>(a,b)</td>
</tr>
<tr>
<td>Hederagenin</td>
<td>(a,b)</td>
</tr>
<tr>
<td>Medicagoic acid</td>
<td>(a,b)</td>
</tr>
<tr>
<td>Soyasapogenol B</td>
<td>(a,b)</td>
</tr>
<tr>
<td>Soyasapogenol E</td>
<td>(a,b)</td>
</tr>
<tr>
<td>Zhanic acid</td>
<td>(b)</td>
</tr>
<tr>
<td>3-Glc-Ara, 28-Glc hederagenin</td>
<td>(a,b)</td>
</tr>
<tr>
<td>3-Glc, 28-Ara-Rha-Xyl medicagenic acid</td>
<td>(a,b)</td>
</tr>
<tr>
<td>3-Glc, 28-Glc-malonyl medicagenic acid</td>
<td>(a,b)</td>
</tr>
<tr>
<td>3-Glc-Glc medicagenic acid</td>
<td>(a,b)</td>
</tr>
<tr>
<td>3-Glc-Glc-Rha, 28-Glc medicagenic acid</td>
<td>(a,b)</td>
</tr>
<tr>
<td>3-Rha-Gal-Glc soyasapogenol B</td>
<td>(a,b)</td>
</tr>
<tr>
<td>3-Glc-Glc-Glc, 23-Ara-28-Ara-Rha-Xyl zhanic acid</td>
<td>(b)</td>
</tr>
<tr>
<td>Afrormosin</td>
<td>(c)</td>
</tr>
<tr>
<td>Afrormosin 7-O-α-glucoside</td>
<td>(c)</td>
</tr>
<tr>
<td>Afrormosin 7-O-α-glucoside-6'-O-malonate</td>
<td>(c)</td>
</tr>
<tr>
<td>Alftalone</td>
<td>(d)</td>
</tr>
<tr>
<td>6-Hydroxyflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>Luteolin 7-O-α-glucoside</td>
<td>(c)</td>
</tr>
<tr>
<td>Biochanin A 7-O-α-glucoside</td>
<td>(c)</td>
</tr>
<tr>
<td>Biochanin A 7-O-α-glucoside-6'-O-malonate</td>
<td>(c)</td>
</tr>
<tr>
<td>Daidzein 7-O-α-glucoside</td>
<td>(c)</td>
</tr>
<tr>
<td>Daidzein 7-O-α-glucoside-6'-O-malonate</td>
<td>(c)</td>
</tr>
<tr>
<td>4',2,7-Trihydroxyisoflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>4'-Methoxy, 2,6,7-trihydroxyisoflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>4'-Methoxy, 2,7-dihydroxyisoflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>4'-Methoxy, 6,7-dihydroxyisoflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>7-Methoxy, 4',5,6,7-tetrahydroxyisoflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>6-Methoxy, 4',5,6,7-tetrahydroxyisoflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>4',2,6,7-Tetrahydroxyisoflavone</td>
<td>(d)</td>
</tr>
<tr>
<td>Irlolone</td>
<td>(c)</td>
</tr>
<tr>
<td>Irlolone 4'-O-α-glucoside</td>
<td>(c)</td>
</tr>
<tr>
<td>Irlolone 4'-O-α-glucoside-6'-O-malonate</td>
<td>(c)</td>
</tr>
<tr>
<td>Irsololidone</td>
<td>(c)</td>
</tr>
<tr>
<td>Irsololidone 7-O-α-glucoside</td>
<td>(c)</td>
</tr>
<tr>
<td>Irsololidone 7-O-α-glucoside-6'-O-malonate</td>
<td>(c)</td>
</tr>
<tr>
<td>(--)Medicarpin 3-O-α-glucoside</td>
<td>(c)</td>
</tr>
<tr>
<td>(--)Medicarpin 3-O-α-glucoside-6'-O-malonate</td>
<td>(c)</td>
</tr>
<tr>
<td>Sativan</td>
<td>(d)</td>
</tr>
<tr>
<td>Vestitol</td>
<td>(d)</td>
</tr>
<tr>
<td>Beta-amyrin</td>
<td>(e)</td>
</tr>
</tbody>
</table>

(a) Huhman and Sumner, 2002; (b) Huhman et al., 2005; (c) Farag et al., 2007; (d) Deavours et al., 2006; (e) Broeckling et al., 2005.

### Table 2. Overview of the MedicCyc database including the number of pathways, reactions and compounds

<table>
<thead>
<tr>
<th></th>
<th>MedicCyc version beta</th>
<th>MedicCyc version 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathways</td>
<td>408</td>
<td>254</td>
</tr>
<tr>
<td>Enzymes</td>
<td>3418</td>
<td>3426</td>
</tr>
<tr>
<td>Compounds</td>
<td>1232</td>
<td>1215</td>
</tr>
<tr>
<td>Compounds added since</td>
<td>–</td>
<td>42</td>
</tr>
<tr>
<td>initial automated build</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathways deleted from</td>
<td>–</td>
<td>126</td>
</tr>
<tr>
<td>initial automated build</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathways manually modified since initial automated build</td>
<td>–</td>
<td>17</td>
</tr>
<tr>
<td>Pathways added since initial automated build</td>
<td>–</td>
<td>28</td>
</tr>
</tbody>
</table>

3 RESULTS

3.1 MedicCyc statistics

After completing the initial automated construction of MedicCyc (MedicCyc ver. beta), PathoLogic recognized 3418 enzymes for which the defined functions were known, and another 707 gene products identified as putative enzymes. The putative enzymes consisted of generic annotations such as ‘MAP kinase’ and the precise functions of these enzymes are still unknown. In total, the initial MedicCyc ver. beta database contained 408 pathways (Table 2), which were comprised of 1678 enzymatic reactions and 1232 compounds. It was anticipated that the automated construction would include non-plant pathways because the MetaCyc reference database includes a large number of non-plant species. Accordingly, 126 pathways that could not be confirmed in plants were deleted from the first Pathways Tools analysis.

Seventeen pathways were manually reconstructed (Table 3) to compensate for incomplete pathways retrieved from MetaCyc e.g. superpathway of phenylalanine, tyrosine and tryptophan biosynthesis I. Several pathway fragments were observed in the beta version of MedicCyc due to the lack of Medicago annotations. These included mainly primary metabolic pathways, which are presumed to be conserved across plant species based upon comparisons of the predicted Medicago pathways with pathways from Arabidopsis and rice, i.e. AraCyc and RiceCyc, respectively. Accordingly, the partially reconstructed Medicago pathways noted in Table 3 were extended by copying complete pathways directly from MetaCyc. A few selected pathways related to secondary metabolism were altered using specific knowledge about

### References

- Huhman and Sumner, 2002;
- Huhman et al., 2005;
- Farag et al., 2007;
- Deavours et al., 2006;
- Broeckling et al., 2005.
Medicago metabolism or based upon additional data supporting the presence of novel pathways in Medicago, e.g. detection and identification of the pathway metabolites. Specific new pathway variants included pathways related to isoflavonoid biosynthesis, lignin biosynthesis or anthocyanin biosynthesis.

The final, curated version of MediCyc (1.0) consists of 254 pathways (Table 2), 3426 enzymes and 1215 compounds. Most of the reconstructed pathways are grouped into biosynthesis classes, which are similar to Arabidopsis and rice (Table 4).

The number of enzymatic reactions in different MediCyc pathways varied. However, a different number of Medicago annotated genes were automatically implemented into single pathways, depending upon their biosynthesis class. The most complete reconstructed pathways belong to Energy metabolism, which is analogous to the situation described for Arabidopsis (Mueller et al., 2003). Other pathways such as fatty acids and lipids biosynthesis or amines and polyamines degradation contain only single Medicago genes (e.g. cutin biosynthesis, choline biosynthesis or ureide degradation). It is anticipated that further genomic sequencing and metabolic profiling experiments will provide additional information for verifying MediCyc completeness as we probe deeper into the metabolome of M. truncatula and as more sensitive detection techniques become available.

Biochemical pathway reconstruction provides not only information about predicted pathways, but also about selected reactions for which no corresponding enzymes have been assigned yet. These reactions are referred to as ‘pathway holes’, and are detected by the Pathway Tools software (Keseler et al., 2005). So far 587 pathways holes were identified in MediCyc.

---

**Table 3.** Systematic listing of the manually curated and added pathways in MediCyc database (1.0)

<table>
<thead>
<tr>
<th>Modified pathways</th>
<th>Added pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine degradation I</td>
<td>Abscisic acid biosynthesis</td>
</tr>
<tr>
<td>Arginine biosynthesis I</td>
<td>Afrormosin conjugates interconversion</td>
</tr>
<tr>
<td>Arginine degradation I</td>
<td>Alanine biosynthesis II</td>
</tr>
<tr>
<td>Arginine degradation II</td>
<td>Alfalone and afrormosin biosynthesis</td>
</tr>
<tr>
<td>Choline biosynthesis I</td>
<td>Biochanin A conjugates interconversion</td>
</tr>
<tr>
<td>dTDP-l-rhamnose biosynthesis II</td>
<td>Chlorophyll cycle</td>
</tr>
<tr>
<td>Fatty acid biosynthesis — initial steps</td>
<td>Cutin biosynthesis</td>
</tr>
<tr>
<td>Formononetin and derivatives biosynthesis</td>
<td>Cytokinins degradations</td>
</tr>
<tr>
<td>Glycosyglyceride biosynthesis</td>
<td>Daizdein conjugates interconversion</td>
</tr>
<tr>
<td>Isoleucine degradation I</td>
<td>Formononetin conjugates interconversion</td>
</tr>
<tr>
<td>Lignin biosynthesis</td>
<td>Formononetin, medicarpin and sativan biosynthesis</td>
</tr>
<tr>
<td>Proline biosynthesis</td>
<td>Gibberellin inactivation</td>
</tr>
<tr>
<td>Superpathway of citrulline metabolism</td>
<td>Irlone biosynthesis</td>
</tr>
<tr>
<td>Superpathway of fatty acid biosynthesis</td>
<td>Irlone conjugates interconversion</td>
</tr>
<tr>
<td>Superpathway of phenylalanine, tyrosine and tryptophan biosynthesis I</td>
<td>Irisolidone conjugates interconversion</td>
</tr>
<tr>
<td>Superpathway of sulfur amino acid biosynthesis</td>
<td>Luteolin conjugates interconversion</td>
</tr>
<tr>
<td>Valine degradation I</td>
<td>Mannitol biosynthesis</td>
</tr>
<tr>
<td></td>
<td>Medicarpin conjugates interconversion</td>
</tr>
<tr>
<td></td>
<td>NAD biosynthesis II (from tryptophane)</td>
</tr>
<tr>
<td></td>
<td>NAD salvage pathway II</td>
</tr>
<tr>
<td></td>
<td>Pantothenate biosynthesis</td>
</tr>
<tr>
<td></td>
<td>Phenylalanine biosynthesis</td>
</tr>
<tr>
<td></td>
<td>Salicylic acid biosynthesis</td>
</tr>
<tr>
<td></td>
<td>Spermine biosynthesis I</td>
</tr>
<tr>
<td></td>
<td>Triterpene saponines biosynthesis</td>
</tr>
<tr>
<td></td>
<td>Tyrosine biosynthesis I</td>
</tr>
<tr>
<td></td>
<td>Tyrosine biosynthesis II</td>
</tr>
<tr>
<td></td>
<td>UDP-l-rhamnose biosynthesis</td>
</tr>
</tbody>
</table>

**Table 4.** Summary of the pathways in MediCyc 1.0 grouped by category

<table>
<thead>
<tr>
<th>Categories</th>
<th>No. of pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biosynthesis</td>
<td>191</td>
</tr>
<tr>
<td>Amines and polyamines</td>
<td>10</td>
</tr>
<tr>
<td>Amino acids</td>
<td>56</td>
</tr>
<tr>
<td>Aminoacyl-tRNAs</td>
<td>1</td>
</tr>
<tr>
<td>Aromatic compounds</td>
<td>1</td>
</tr>
<tr>
<td>Cell structure</td>
<td>9</td>
</tr>
<tr>
<td>Cofactors, prosthetic groups, electron donors</td>
<td>35</td>
</tr>
<tr>
<td>Fatty acid and lipids</td>
<td>16</td>
</tr>
<tr>
<td>Hormones</td>
<td>12</td>
</tr>
<tr>
<td>Nucleosides and nucleotides</td>
<td>5</td>
</tr>
<tr>
<td>Secondary metabolism</td>
<td>40</td>
</tr>
<tr>
<td>Sugar and polysaccharides</td>
<td>14</td>
</tr>
<tr>
<td>Others</td>
<td>2</td>
</tr>
<tr>
<td>Degradation</td>
<td>72</td>
</tr>
<tr>
<td>Alcohols</td>
<td>1</td>
</tr>
<tr>
<td>Amines and polyamines</td>
<td>2</td>
</tr>
<tr>
<td>Amino acids</td>
<td>22</td>
</tr>
<tr>
<td>CI compounds</td>
<td>1</td>
</tr>
<tr>
<td>Fatty acid and lipids</td>
<td>8</td>
</tr>
<tr>
<td>Hormones</td>
<td>2</td>
</tr>
<tr>
<td>Inorganic nutrients</td>
<td>6</td>
</tr>
<tr>
<td>Nucleosides and nucleotides</td>
<td>2</td>
</tr>
<tr>
<td>Secondary metabolism</td>
<td>3</td>
</tr>
<tr>
<td>Sugar derivatives</td>
<td>7</td>
</tr>
<tr>
<td>Sugar and polysaccharides</td>
<td>14</td>
</tr>
<tr>
<td>Others</td>
<td>5</td>
</tr>
<tr>
<td>Generation of precursor metabolites and energy</td>
<td>13</td>
</tr>
<tr>
<td>Pentose phosphate pathways</td>
<td>3</td>
</tr>
<tr>
<td>Photosynthesis</td>
<td>3</td>
</tr>
<tr>
<td>Respiration</td>
<td>2</td>
</tr>
<tr>
<td>TCA cycle</td>
<td>3</td>
</tr>
<tr>
<td>Super-pathways</td>
<td>31</td>
</tr>
</tbody>
</table>
It is presumed that a significant portion of the missing enzymes will be revealed as the *M. truncatula* genomic sequencing project moves closer towards completion.

### 3.2 Functional genomics data visualization

It is useful to visualize functional genomics data within a biochemical context for the interpretation of large-scale ‘omics’ data. Thus, reconstructed Medicago pathways can also be used in data analysis via the ‘Omics Viewer’, a software module of Pathways Tools. The ‘Omics Viewer’—can visualize data from gene expression, protein expression, gene family analysis or metabolite profiling experiments within the biological context of metabolic networks (Zhang et al., 2005). Each reaction (represented as a connection line between the compounds) can be color coded in accordance with the expression level of the gene and the protein catalyzing the reaction (Fig. 1). As we used expressed sequence information from the Medicago Affymetrix Chipset, MedicCyc was engineered to visualize experimental transcript data specifically from this platform. The results obtained using alternative microarray platforms can also be accommodated assuming each transcript element is represented with a unique nomenclature (TIGR ID or IMGAG ID). The MedicCyc Omics Viewer was also designed to display metabolic profiling data. Each detected metabolite, which was previously implemented into MedicCyc database, can be color coded in accordance with its recorded abundance. Correlated metabolite and expression data can then be viewed in the form of a general overview map (Fig. 1A) or as in the expanded single pathway view (Fig. 1B and C). The visualization viewer is available through the MedicCyc web site which enables visualization of user’s own data.

### 4 DISCUSSION AND CONCLUSION

The aims of this project were to use pathway reconstruction to predict the metabolic composition of *M. truncatula* and further exploit this information as a platform for the visualization of integrated functional genomics datasets. This was achieved through the construction of MedicCyc. This database was manually curated and the final curated version of MedicCyc (1.0) consists of 254 pathways, 3426 enzymes and 1215 compounds (Table 2). A significant number of enzymes were identified as still missing from the currently incomplete *M. truncatula* genome sequencing project through comparisons with AraCyc. As the *M. truncatula* genome sequencing project moves toward completion, MedicCyc will be updated. The MedicCyc database is currently supported by The Noble Foundation and additional funding for continued curation and improvements will be sought. Continued curation will be necessary to address new transcript annotations and new metabolic pathways related to Medicago metabolism which become available during on-going sequencing and annotation of the *M. truncatula* genome. Additional future efforts will focus on the inclusion of subcellular localizations for specific known enzymatic isoforms and metabolites. Continual metabolic profiling experiments are expected to provide additional information concerning novel metabolic content in *M. truncatula* and will be added as available. It is anticipated that this resource will serve as a repository for our current and future understanding of Medicago metabolism, provide a fundamental tool for the visualization of functional genomics datasets, serve as a contributor to larger databases (MetaCyc, MtDB, LIS, etc.), facilitate comparative studies of pathways across species, enable the prediction and annotation of
unknown genes, and facilitate metabolic engineering of legume metabolism for the enhancement of desired traits. We strongly encourage input from the scientific community through the incorporated tool available online.

ACKNOWLEDGEMENTS

We thank Ms Keisha Wood for assistance with the database curation and Mr Roland Stolfi for technical assistance in establishing the MedicCyc database. We also thank Drs Patrick Zao and Xinbin Dai for consultation, and we express our appreciation to Dr Yuhong Tang for providing unpublished Affymetrix microarray data. This work was financially supported by The Samuel Roberts Noble Foundation. Funding to pay the Open Access publication charges was provided by The Samuel Roberts Noble Foundation.

Conflict of Interest: none declared.

REFERENCES


