EGENES: Transcriptome-Based Plant Database of Genes with Metabolic Pathway Information and Expressed Sequence Tag Indices in KEGG\(^1\)[C]\(^W\)\[^OA\]

Ali Masoudi-Nejad*, Susumu Goto, Ruy Jauregui\(^2\), Masumi Ito, Shuichi Kawashima, Yuki Moriya, Takashi R. Endo, and Minoru Kanehisa

Laboratory of Bioknowledge Systems, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho Uji, Kyoto 611–0011, Japan (A.M.-N.); Laboratory of Bioinformatics and Bioknowledge Systems, Institute of Biochemistry and Biophysics, and Center of Excellence for BioMathematics, School of Computer Science, University of Tehran, Tehran, Iran (A.M.-N.); Laboratory of Genome Database, Human Genome Center, University of Tokyo, Tokyo 108–8639, Japan (S.K., M.K.); and Laboratory of Plant Genetics, Division of Applied Bioscience, Kyoto University, Gokasho Uji, Kyoto 611–0011, Japan (T.R.E.)

EGENES is a knowledge-based database for efficient analysis of plant expressed sequence tags (ESTs) that was recently added to the KEGG suite of databases. It links plant genomic information with higher order functional information in a single database. It also provides gene indices for each genome. The genomic information in EGENES is a collection of EST contigs constructed from assembly of ESTs. Due to the extremely large genomes of plant species, the bulk collection of data such as ESTs is a quick way to capture a complete repertoire of genes expressed in an organism. Using ESTs for reconstructing metabolic pathways is a new expansion in KEGG and provides researchers with a new resource for species in which only EST sequences are available. Functional annotation in EGENES is a process of linking a set of genes/transcripts in each genome with a network of interacting molecules in the cell. EGENES is a multispecies, integrated resource consisting of genomic, chemical, and network information containing a complete set of building blocks (genes and molecules) and wiring diagrams (biological pathways) to represent cellular functions. Using EGENES, genome-based pathway annotation and EST-based annotation can now be compared and mutually validated. The ultimate goals of EGENES will be to: bring new plant species into KEGG by clustering and annotating ESTs; abstract knowledge and principles from large-scale plant EST data; and improve computational prediction of systems of higher complexity. EGENES will be updated at least once a year. EGENES is publicly available and is accessible by the following link or by KEGG’s navigation system (http://www.genome.jp/kegg-bin/create_kegg_menu?category=plants_egenes).

In the past decade, bioinformatics has become an integral part of research and development in the biological sciences. Bioinformatics now has an essential role both in deciphering genomic, transcriptomic, and proteomic data generated by high-throughput experimental technologies and in organizing information gathered from traditional biology. Sequence-based protocols for analyzing individual genes or proteins have been elaborated and expanded, and different methods have been developed for analyzing large numbers of genes or proteins simultaneously. To date, the genome sequence for over 349 different species has been published, and sequencing of 987 other prokaryotic and 588 eukaryotic genomes is under development (Liolios et al., 2006; http://www.genomesonline.org). However, owing to the cost of sequencing and their relatively large size, the full genome sequences of only a few plants among 41 published eukaryotes are currently available. Automatic sequencing has had an enormous impact on the high-throughput generation of various biological data such as ESTs and single nucleotide polymorphisms. Due to the extremely large genome, the mass collection of data containing single-pass (low-quality and fragment) sequences is a quick way to capture a complete repertoire of genes expressed in specific cells or tissues. On the other hand, an increasing number of projects have been initiated to generate survey sequence data for a large number of plants. Such data consist of many thousands of short (usually approximately 300–500 bp), single-pass sequence reads...
from either mRNA/cDNA (EST) or genomic DNA (genome survey sequences [GSSs]). There are currently over 1,000 species across a variety of taxonomic groups for which 29,425,525 ESTs and 12,351,453 GSSs have been generated (National Center for Biotechnology Information [NCBI]/dbEST, release 092305, September 2005) in total. In general, the poorly organized nature of these data makes them difficult to interpret within a genomic context and precludes even simple comparative analyses. Common problems include significant redundancy in the datasets, vector/adaptor contamination, poor quality, and a lack of consistent annotation between projects. An effective way to overcome these problems is to group ESTs and GSSs into clusters (representing unique genes), which may be subsequently fed into downstream annotation pipelines. Several well-known EST clustering algorithms exist, most of which depend on pairwise alignment of ESTs. NCBI’s UniGene (Schuler et al., 1996; http://www.ncbi.nlm.nih.gov/UniGene) is the most widely used database created using such algorithm. The Institute for Genomic Research (TIGR) Gene Indices is another well-known EST clustering procedure that combines EST clustering based on sequence similarity (Lee et al., 2005; http://www.tigr.org/tdb/tgi). STACKdb also combines EST clustering with transcript assembly and is designed to examine transcript variations in the context of developmental and pathological states (Christoffels et al., 2001). Recently, many other algorithms have been introduced for EST clustering that must be thoroughly tested before using them for large-scale projects (Kim et al., 2005; Ptitsyn and Hide, 2005; Schneeberger et al., 2005). There are databases that describe genes, their protein products, and the biological processes in which they are involved. The purpose of these databases is to unify the biological datasets specific to a species or family at a single resource (Ware et al., 2002; Carollo et al., 2005; Kunne et al., 2005; Schneider et al., 2005; Wheeler et al., 2005; Kurata and Yamazaki, 2006).

To increase our understanding of cellular processes from genome information, pathway databases such as KEGG (Kanehisa, 1997; Kanehisa et al., 2006) and EcoCyc (Karp et al., 1998) have been created in the past decade. Whereas most databases concentrate on molecular properties, these databases tackle complex properties of cellular pathways, such as metabolism, signal transduction, and the cell cycle, by storing the corresponding networks of interacting molecules in computerized forms, often as graphical pathway diagrams. Inevitably, it is necessary to collect data and knowledge from published literature accumulated over many years from traditional studies of biology. At least for metabolic pathways, the past knowledge is relatively well organized in these databases, providing a reference data set for annotating genomes (i.e. metabolic reconstruction) and for screening microarray and other high-throughput experimental data. Although the Internet-based linking among different databases is a convenient way to use this huge resource, more effort is required for the true integration of biological knowledge and the genome information.

EGENES is a new database recently added to the KEGG suite of databases (Kanehisa et al., 2006). EGENES (http://www.genome.jp/kegg-bin/create_kegg_menu?category=plants_egenes) is a knowledge-based system for efficient analysis of plant ESTs, linking genomic information with higher order functional information in a single database. The genomic information is provided as a collection of EST contigs, produced from assembling the public ESTs (Masoudi-Nejad et al., 2004, 2006), with a gene/EST index such as functional annotation and source information for each genome. The higher order functional information is stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, genetic information processing, environmental information processing, and cellular process. Functional assignments is a process of linking a set of genes/transcripts in each genome with a network of interacting molecules in the cell, such as a pathway or a complex, representing a higher order biological function. EGENES follows the same protocols for functional assignment as other KEGG resources.

In this article, we provide an introduction to EGENES and discuss its importance for the plant research community. Because all the resources in KEGG follow the same architecture and design, an appraisal of EGENES should give readers an idea of the available information stored in KEGG and how to use it efficiently.
by combining and assembling repetitive elements using Phrap and CAP3 and assembling them into a single database. EGrep contains 214,169 repeat-related sequences assembled into 3,206 contigs and 3,573 singletons.

The pipeline used for processing, assembling, and KEGG-based annotation of the ESTs is summarized in Figure 1 and was described previously (Masoudi-Nejad et al., 2004, 2006; Moriya et al., 2005). The pipeline includes: sequence cleaning, repeat masking, vector masking, sequence assembling, and KEGG annotation.

The sequence cleaning process involves basic procedures, such as removing the polyA/polyT tail, clipping low-quality ends (the ends rich in undetermined bases), and discarding those that are too short (shorter than 100) or mostly low-complexity sequences. The repeat masking process uses the RepeatMasker program (Smit et al., 2004) that compares the query sequence against our custom repeat database (EGrep). Vector masking is performed using the program Cross_Match (Ewing et al., 1998), which is used to compare query sequences to our custom vector database EGvec and produce vector masked versions of the input sequences. The sequence assembling process uses the CAP3 program (Huang and Madan, 1999) for clustering and assembling the sequences into contigs and singletons. CAP3 assembles ESTs from the same gene under more stringent criteria than the other approaches and was shown to be superior to TIGR assembler (Liang et al., 2000) and Phrap (http://www.phrap.org/phredphrap/phrap.html) in its ability to distinguish gene family members while tolerating sequencing error. Because the efficiency of CAP3 can also be improved or changed by incorporating different options in the clustering algorithm, we tested CAP3 to cluster ESTs with various stringency criteria (sequence identity $P = 80\%$, $90\%$, $92\%$, $95\%$, and $97.5\%$). Results showed that $P = 92\%$ is the best threshold, as it is stringent enough to separate paralogs while being capable of tolerating sequencing error to avoid misclassification of ESTs from the same gene into two or more clusters. A range of overlap lengths was initially examined, but the results showed no significant difference, so an overlap length of 40 (default) was used for clustering. Finally, the data were processed with CAP3 default options and with $P = 92$. Results of clustering and assembling (only EST contigs and not singletons) were chosen for automatic annotation and mapping to the KEGG pathway.

For annotation, EST contigs were compared against each genome in the KEGG GENES database by BLAST searches (Moriya et al., 2005) in both forward (BLASTX) and reverse directions (TBLASTN), taking each gene in genome $A$ as a query compared against all genes in genome $B$, and vice versa. Those BLAST hits with bit scores less than 60 were removed. Because the bit scores of a gene pair $a$ and $b$ from two genomes $A$ and $B$, respectively, can be different in forward and reverse directions and because the top scores do not necessarily reflect the order of the rigorous Smith-Waterman scores, we used bidirectional hit rate (BHR) and select genes with BHR greater than 0.95 (Moriya et al., 2005). Next, we classified the ortholog candidates into groups by their KEGG orthology (KO) numbers using the KEGG Automatic Annotation Server (http://www.genome.jp/kegg/kaas/) that assigns KO identifiers as controlled vocabularies to sets of new sequences based on BHR. The KO identifier, or the KO number, is a manually curated functional classification based on the KEGG PATHWAY database and BRITE functional hierarchy (Kanehisa et al., 2004) and is a common identifier for linking the genomic information in the KEGG GENES database and the network information in the KEGG PATHWAY database. The pathway nodes represented by rectangles in the KEGG reference pathway maps are given KO identifiers, so that organism-specific pathways are automatically generated once each genome is annotated with KOs. In this study, organism-specific pathway maps were computationally generated for 25 plants if at least two enzymes were present in the corresponding reference maps based on the KO assignments. EGENES also provides a gene index for each plant species. For each EST contig, the database provides a FASTA text-alignment file including all the EST members of the contigs and a graphical alignment diagram with links to the original GenBank entries for ESTs included in the alignment. The EST contigs are the result of stringent clustering and assembling, which produces a set of unique and virtual transcripts. These contig sequences can be used as a resource for comparative and functional genomic analysis by providing putative genes with functional annotation.

CONTENT OF EGENES

The current version of EGENES (release 41.0, January 2007) consists of 25 species. Table I shows the species included in the initial version of EGENES. Two other genome-based plant entries for Arabidopsis (Arabidopsis thaliana) and rice (Oryza sativa) are also available in KEGG (the genome-based pathways were produced by manual KO annotation based on Smith-Waterman sequence similarity search of the TIGR gene annotation of latest release of the Arabidopsis and rice to the KEGG GENES database). The table shows corresponding KEGG codes covering 11 families (mustard, rue, mallow, pea, willow, daisy, nightshade, grape, grasses, pine, and moss). Entries with ESTs have four-character codes that start with ‘e’ (e.g., ‘eath’ for the Arabidopsis ESTs data), and complete genome sequences have three-character codes (‘ath’ represents the Arabidopsis genome data). By clicking on the species code in the EGENES entry page, a new window provides relevant information for each species, including taxonomy data (Fig. 2A). The data for each species in EGENES are organized in two distinct parts: pathway maps and gene catalogs.
Figure 1. Flow of data processing in EGENES. The flowchart shows the pipeline used in EGENES. For details of each step, see the text.
PATHWAY MAPS

The pathway maps for each species were derived from KEGG’s reference pathways based on the automatic annotation as described above. The KEGG PATHWAY database contains 266 reference pathways, all of which have been manually curated. Each reference pathway can be viewed as a network of enzymes or a network of enzyme commission (EC) numbers. Knowledge-based prediction of metabolic pathways involves the matching of genes in the genome against enzymes in the KEGG reference pathway.

Historically, the integration of pathway information and genomic information was first achieved in KEGG by EC numbers. The EC system uses four numbers that define in a hierarchical manner the function of an enzyme. However, there can be different enzymes that have the same EC number but function in different pathways. This means that, in cases where EC numbers refer to more than one enzyme, we have to determine which enzyme is appropriate to a particular KEGG pathway. However, to incorporate nonmetabolic pathways and to overcome various problems inherent in the enzyme nomenclature, a new scheme based on the ortholog identification (KO) was introduced, replacing the EC numbers. KO is based on computational analysis, as well as manual curation, decomposing all genes in the complete genomes into sets of orthologs. Here, two genes are considered as orthologs, or belonging to the same KO group, when they are mapped to the same KEGG pathway node.

KO will be used not only to characterize all known pathways but also to explore unknown pathways that have not been experimentally verified but can be inferred by sequence similarity to enzymes from other species. Once enzyme genes are identified in the genome based on sequence similarity and the EC numbers and the KO identifiers are properly assigned, species-specific pathways could be constructed computationally by correlating EST contigs with gene products (enzymes) in the reference pathways according to the matching KO identifiers and EC numbers.

Figure 2B shows a reference pathway diagram in KEGG. Each pathway consists of rectangles representing enzymes labeled with their EC numbers, small circles representing reactants of the enzyme reactions, and their connections. Here, each rectangle has the corresponding KOs as underlying function information. Because the KO can be used for linking enzymes and genes coding for the enzymes in each species, this graphic representation is most useful in superimposing the genomic information onto the knowledge of metabolic pathways, which helps to deduce metabolism for each species. Because the metabolic pathway, especially for intermediary metabolism, is well conserved among most species and we have already annotated enzyme genes from completely sequenced genomes, it is possible to manually draw one reference pathway and consequently computationally generate many species-specific pathways by using EGENES annotation based on BHR.

Table 1. Species included in current version of the EGENES database (release 40.0, December 2006)

| Family       | Species (Common Name) | Code | EST | Contigs | Singletons | Pathway Maps | Mapped Contigs |
|--------------|-----------------------|------|-----|---------|------------|--------------|---------------|
| Mustard      | Arabidopsis (thale cress) | eath | 322,641 | 23,297 | 57,348 | 118 | 1,235 |
|             | Brassica napus (rape) | ebna | 37,159 | 5,602 | 10,081 | 92 | 890 |
| Rue          | Citrus sinensis (Valencia orange) | ecsi | 45,192 | 5,687 | 8,491 | 107 | 737 |
| Mallow       | Gossypium arboreum (cotton) | egar | 39,097 | 5,494 | 15,052 | 96 | 732 |
| Gossypium raimondii |egra | 67,577 | 8,764 | 16,933 | 117 | 932 |
| Pea          | Glycine max (soybean) | egma | 334,036 | 30,291 | 51,791 | 122 | 1,598 |
|             | Lotus corniculatus (lotus) | elco | 111,455 | 12,731 | 19,679 | 108 | 776 |
|             | Medicago truncatula (barrel medic) | emtr | 187,763 | 17,456 | 21,467 | 121 | 1,252 |
| Willow       | Populus tremula (aspen) | eptp | 31,288 | 4,824 | 9,518 | – | – |
|             | Populus balsamifera (poplar) | ephb | 54,660 | 7,595 | 11,967 | 103 | 872 |
| Daisy        | Helianthus annuus (sunflower) | ehan | 59,841 | 6,346 | 15,446 | 97 | 727 |
|             | Lactuca sativa (lettuce) | elsa | 68,188 | 8,114 | 15,365 | 99 | 764 |
| Nightshade   | Lycopersicon esculentum (tomato) | eles | 150,596 | 15,397 | 16,513 | 116 | 1,244 |
|             | Solanum tuberosum (potato) | estu | 158,162 | 19,422 | 19,327 | 122 | 1,326 |
| Grape        | Vitis vinifera (wine grape) | eivi | 141,616 | 13,575 | 12,774 | 113 | 939 |
| Grass        | Rice | eosa | 284,007 | 23,217 | 61,434 | 120 | 1,283 |
|             | Hordeum vulgare (barley) | ehvu | 381,006 | 23,021 | 57,465 | 119 | 1,332 |
|             | Triticum aestivum (wheat) | etae | 558,098 | 39,461 | 48,458 | 118 | 1,815 |
|             | Zea mays (maize) | ezma | 415,211 | 30,113 | 56,281 | 119 | 1,527 |
|             | Saccharum officinarum (sugarcane) | esof | 246,301 | 19,597 | 23,171 | 122 | 1,189 |
|             | Sorghum bicolor (sorghum) | esbi | 190,946 | 19,422 | 19,327 | 122 | 1,326 |
| Pine         | Pinus taeda (loblolly pine) | epta | 173,680 | 15,397 | 15,334 | 113 | 1,101 |
|             | Picea glauca (white spruce) | eppp | 102,334 | 13,249 | 11,453 | 109 | 1,015 |
|             | Phycomitrella patens | eppp | 82,420 | 12,105 | 9,127 | – | – |

Copyright © 2007 American Society of Plant Biologists. All rights reserved.
Pathway maps have been hierarchically organized into four main categories, each of which consists of many subcategories (Kanehisa et al., 2006). Each subcategory is in fact a collection of pathway diagrams. Table II shows a summary of the pathway in EGENES. In total, 178 unique diagrams were created for all plants, organized into 134 in Metabolism, 19 in Genetic Information Processing, 21 in Environmental Information Processing, and four in Cellular Processes.

**Exploring the Pathway Maps**

By clicking the pathway maps link in Figure 2A, a hierarchical context window will appear, displaying all the pathways present in the current species (Supplemental Fig. S1). Each number to the left of the pathway map name is the pathway map identification. Clicking on each pathway will show the diagram for that pathway, including all enzymes and compounds in the reference pathway. For example, by clicking on Photosynthesis under Energy Metabolism, the corresponding pathway diagram will be shown (http://www.genome.jp/dbget-bin/get_pathway?org_name=eath&mapno=00195). White boxes (e.g. PsbB, PsbF, etc.) indicate that there are no Arabidopsis ESTs similar to those genes/enzymes in other species. Colored boxes are the enzymes that are present in Arabidopsis (EC 1.18.1.2, PetB, PsbA, PsbO, etc.). By clicking on any colored box, gene catalog data for that gene (EST) are shown.

Figure 2. Typical information stored in KEGG. A, Entry point for each species in the EGENES database, including pathway maps, gene catalogs, taxonomy information, and source sequences used for sequence assembly and analysis. B, A typical pathway diagram in KEGG (reductive carboxylate cycle). C, Information corresponding to KO K00844 and the complete information for this contig in different pathways. D, Graphical view of the alignment of the members of contig 3,213. [See online article for color version of this figure.]
GENE CATALOGS

By clicking the Gene Catalogs link in Figure 2A, the entry point for the EST-based gene catalog of Arabidopsis will appear (http://www.genome.jp/kegg-bin/show_organism?menu_type=genecatalogs&org=eath). The gene catalog is the EST/gene index and EGENES functional classification, which includes functional hierarchies of KEGG pathways, ortholog groups (KO), and protein families. Clicking on KO will show the first level of the pathway-based classification of ortholog groups. The first level contains Metabolism, Genetic Information Processing, Environmental Information Processing, and Cellular Processes (http://www.genome.jp/dbget-bin/get_htext?eath00001.keg+p+/-kegg/brite/eath+f+F+F). Second and third levels show further hierarchies inside each part of the first level. The fourth level shows the list of contigs orthologous to each enzyme in each category (http://www.genome.jp/dbget-bin/get_htext?eath00001.keg+p+/-kegg/brite/eath+f+F+F.D). For example, the second contig that appears is 3,213, corresponding to KO K00844.

By clicking on K00844, a new window shows the information related to the contig including other related pathways (Fig. 2C). An ortholog of this contig is hexokinase (EC 2.7.1.1), which is present in glycolysis (PATH Ko00010), Fru and Man metabolism (PATH ko00051), Gal metabolism (PATH Ko00052), starch and Suc metabolism (PATH Ko00500), and aminosugars metabolism (PATH ko00530), all of which belong to carbohydrate. Figure 2C also shows other associated information, such as EC, reaction, substrate, product, pathway, and Gene Ontology (GO) number. Reaction includes the reactions in which this enzyme participates. By clicking on any reaction identifier (e.g. R00867), users will get new windows showing the associated reactions and compounds (http://www.genome.jp/dbget-bin/www_bget?rn+R00867). Clicking on any Compound (e.g. C00002) shows the compound on which this enzyme acts (http://www.genome.jp/dbget-bin/www_bget?compound+C00002).

Clicking on 3213 shows Arabidopsis EST contig 3,213, similar to its corresponding enzyme. The main window has seven sub-windows, including KO annotation (gene catalogs), EST members of the contig with links to graphical and text alignment, links to pathway and other databases, and the contig’s sequence (http://www.genome.jp/dbget-bin/www_bget?ko+k00844). Clicking on the Alignment button shows the text alignment of EST sequences that forms this contig, which in fact is the output of the CAP3 (Huang and Madan, 1999) assembling program (see Supplemental Fig. S1). The graphical view of the alignment is also provided to illustrate how the contig members overlap (Fig 2D). On the left side, the GenBank number (gi) of each single EST is shown. Clicking on it shows the corresponding GenBank data.

PATHWAY COVERAGE

To assess whether the coverage of EST contigs is sufficient for computationally predicting pathways, we checked the number of pathway nodes present in the same pathway between the manually curated pathways for Arabidopsis (ath) and rice (osa) and the predicted pathways for these plants based on ESTs (eath and eosa). In many pathways, we found that EST coverage is more than genomic DNA. In total, we found 251 enzymes (pathway nodes) unique to the EST-based rice metabolic pathway (osa) but not found in the rice pathway based on genomic DNA. For Arabidopsis, we found 295 unique enzymes that are found only in the EST-based predicted pathway. Figure 3, A and B, shows diagrams of genome (ath) and EST-based (eath) inositol phosphate metabolism paths for Arabidopsis, respectively. Total nodes/reactions complete: ath = 7; eath = 14. Reactions common and complete between ath and eath = 3 (2.7.1, 2.7.1.68, 5.5.1.4). Reactions unique to ath = 4 (3.1.3.57 [2×], 3.1.3.67, 2.7.1.137), and reactions unique to eath = 10 (2.7.1.34, 2.7.1.134 [2×], 3.1.3.25 [3×], 3.1.3.36, 2.7.1.67, 4.6.1.13, 1.13.99.1, Supplemental Table S1).

SIMILARITY SEARCH AND PHYLOGENIC ANALYSIS

Another entry point for EGENES is the sequence similarity search in the KEGG database using BLAST (http://blast.genome.jp/) and FASTA (http://fasta.
For example, to find all the sequences similar to Arabidopsis cDNA clone gi 32885696, click on the BLAST button, copy and paste the sequence into the form, choose BLASTN and the database to search, KEGG EGENES (Fig. 3C), and press the compute button. All the similar contigs from EGENES (including species other than Arabidopsis) will be shown. From here, a variety of analyses on the BLAST results is available, such as performing multiple alignments (ClustalW and MAFFT) and graphical alignments of BLAST search and drawing a phylogenic tree (Fig. 3D) via the GenomeNet server (http://align.genome.jp/).

**DISCUSSION AND FUTURE DIRECTIONS**

With the release of the fully sequenced genomes of Arabidopsis and rice and the initiation of sequencing projects for many other plant species and large-scale ESTs sequencing for many other plant species, there is a fast-growing desire to place this genomic information into a metabolic context. ESTs provide researchers with a quick and inexpensive route for discovering new genes, obtaining data on gene expression and regulation, and constructing genome maps. Efforts to computerize our knowledge on cellular functions, at present either by the controlled vocabulary of GO or by graph representation, will facilitate the computational mapping of genomic data to complex cellular properties or detect any empirical relationships between genomic and higher properties.

There are several molecular biology databases available on the Internet, such as AraCyc (Mueller et al., 2003; http://arabidopsis.org/tools/aracyc) and RiceCyc (http://www.gramene.org/pathway/), which are
devoted to plant species, and many more will appear as genomic data and computational methods are introduced into individual research in biology. The benefits of species-specific metabolic pathway databases are substantial and have been thoroughly discussed (Zhang et al., 2005). A species-specific pathway depicts the biochemical components of an organism, assists comparative studies of pathways across species, and facilitates metabolic engineering to improve crop metabolism and traits. Although the manual creation of a plant pathway database is a lofty goal, it is very labor intensive and time consuming because of the extreme complexity of plant genomes and a lack of information. An alternative is to computationally predict species-specific plant pathways as a starting point for later manual curation (Mueller et al., 2003).

Using accurate and comprehensive reference databases is the key to ensuring the quality of derived databases. Examples of comprehensive pathway databases are KEGG and MetaCyc (Krieger et al., 2004). Another example for this is AraCyc, which uses the annotated sequences of the Arabidopsis genome from TIGR and The Arabidopsis Information Resource (TAIR; Mueller et al., 2003) using a GO term (Gene Ontology Consortium, http://www.genetonology.org) and then predicting the pathway based on MetaCyc. Recently, they have manually validated many parts of the pathway (Zhang et al., 2005) using the literature.

Although both KEGG and EcoCyc attempt to be as complete as possible, there are differences between them in terms of combination and coverage. These differences also exist between EGENES and AraCyc. Each database has its unique features. AraCyc uses only Arabidopsis genomic data that is based on annotation by TIGR and TAIR and is highly contingent on the quality of the annotation in the input data. KEGG includes 26 species for plants (EGENES 25) and two Arabidopsis and rice draft genomes. Indeed, KEGG has two entries for Arabidopsis and rice based on genome data (ath and eath) and EST data (eath and eosa). There are inevitably false positive associations of genes with pathways using only genomic data and automatic assignments (Zhang et al., 2005). For example, isoenzymes localized in different subcellular compartments, though catalyzing the same reaction, may be involved in different pathways. These isoenzymes may not be distinguished by automatic prediction and thus may be assigned to pathways in which they are not involved (Zhang et al., 2005). Using genomic and EST data together will help to improve assignments of genes/enzyme to reactions and pathways solely based on predicted gene-based annotation data. Also in EGENES, plant primary and secondary metabolites are both important, and, because KEGG is accumulating secondary metabolism from various species, EGENES has also included it.

The comparison of coverage of the metabolic reconstruction showed that the EST data increased the coverage more than the genomic information in many instances. In some cases, the number of pathway hits was similar but consisted of hits unique to one or the other, such as the citrate cycle (tricarboxylic acid) in rice, where both eosa and eath have 16 hits that are unique to either of them (data not shown). In Arabidopsis, both eath and ath have 22 hits, six of which are uniquely found in either one or the other (data not shown). This may be because either the sequences or annotations of these genomes are still incomplete. In many cases, by BLASTing EST contigs against the genome, we found that in a certain pathway many genes have been missed based on genomic data. Of the many entries in ath and eath, some portions are common to both, and one database has many specific entries not found in the other. These may be caused by the lack of a complete sequence for Arabidopsis and rice (genome assembly lacks the sequenced regions) or by misannotation (many parts annotated as intergenic regions were hit by our EST contigs). There were also cases where genomic-based annotation gave more hits than EST-based annotation. For example, in the pathway for metabolism of photosynthesis, we have 45 hits; 19 hits were unique to ath and 25 hits were common to both ath and eath (data not shown). This is because the numbers of sequenced ESTs, the size of the cDNA library, or the type of tissue where ESTs have been produced is insufficient to cover all pathway nodes.

The EGENES project is an attempt to capture all reactions and pathways that occur in plants based on both genomic and expressed part of the genome. This is, however, only a starting point. We see the role of EGENES as providing a framework of possible reactions that, when combined with expression, genetics and physical location and enzyme kinetic data, provides the infrastructure for the true integration of biological knowledge and data. Integration in this respect does not simply involve methodology, such as links and common interfaces, but rather involves biology. We are working to make EGENES data available in a variety of standard formats, including text and XML. This will also enable data exchange with other pathway databases, such as AraCyc and other molecular interaction databases.

The ultimate integration of biological databases will be a computer representation of living cells and organisms, whereby any aspect of biology can be examined computationally. Our immediate goals for improving the data coverage and quality of EGENES are as follows: increasing the number of plant species included in the EGENES; including other organisms than plants for comparative genomics analysis; improving the EST clustering and assembling algorithm; including the GSS data for clustering; providing the splice-alignment information for ESTs/GSS contigs; manually curating each species pathway by verifying and linking present pathway diagrams and information to literature and adding new pathway diagrams not present in current versions of the KEGG’s reference pathways; improving automatic assignments of genes/ESTs/GSSs to reactions and pathways for avoiding false-positive assignments; linking the ESTs/genes/
GSSs in reactions and pathways to genetic and physical locations in the genome; linking the ESTs/genes/GSSs in reactions and pathway to microarray chip information and tissue/condition information described in dbEST to give further meaningful information, including isoenzyme discrimination that KO does not explicitly handle; and updating EGENES twice a year to cope with the rapid influx of new ESTs resulting from high-throughput sequencing projects.

In summary, EGENES provides overviews of fundamental biological processes in plants in a form that is useful to students working on a single gene/protein, as well as to bioinformaticians striving to make sense of large-scale datasets. EGENES provides biologist-friendly visualization of biological pathway data. EGENES data and pathway information are publicly available for downloading in XML format from KEGG’s FTP site (ftp://ftp.genome.jp/pub/kegg/).

Supplemental Data
The following materials are available in the online version of this article.

Supplemental Figure S1. Supplement to Figure 2A, Hierarchical organization of the pathway in KEGG. Supplement to Figure 2C, The textual alignment of members of the contig 3213.

Supplemental Table S1. Comparison of coverage between manually curated pathways and the predicted pathways for Arabidopsis and rice.

ACKNOWLEDGMENTS
We express our gratitude to Dr. Kiyoko F. Aoki-Kinoshita and Dr. Nelson Hayes for critical reading of the manuscript.

Received December 19, 2006; accepted April 18, 2007; published April 27, 2007.

LITERATURE CITED
Carollo V, Matthews DE, Lazo GR, Blake TK, Hummel DD, Lui N, Hane DL, Anderson OD (2005) GrainGenes 2.0. An improved resource for the small-grains community. Plant Physiol 139: 643–651

Christoffels A, van Gelder A, Greiling G, Miller R, Hide T, Hide W (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nucleic Acids Res 29: 234–238

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using Phred I: accuracy assessment. Genome Res 8: 175–185

Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M (2002) LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 30: 402–404

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 6: 829–845

Kanehisa M (1997) A database for post-genome analysis. Trends Genet 13: 375–376

Kanehisa M, Goto S, Hat弥orI M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354–D357

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–D280

Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M (1998) EcoCyc: encyclopedia of Escherichia coli genes and metabolism. Nucleic Acids Res 26: 50–53

Kim N, Shin S, Lee S (2005) ECgene: genome-based EST clustering and gene modeling for alternative splicing. Genome Res 15: 566–576

Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD (2004) MetaCyc: a multigene database of metabolic pathways and enzymes. Nucleic Acids Res 32: D438–D442

Kunze C, Lange M, Funke T, Miehe H, Tiel T, Grosse I, Schulz U (2005) CR-EST: a resource for crop ESTs. Nucleic Acids Res 33: D619–D621

Kuratana N, Yamazaki Y (2006) ORyzabase. An integrated biological and genome information database for rice. Plant Physiol 140: 12–17

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana A, Antonescu V, Chan A, Cheung P, Quackenbush J (2005) The TIGR Gene Indices: clustering and assembling ESTs and known genes and integration with eukaryotic genomes. Nucleic Acids Res 33: D71–D74

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J (2000) An optimized protocol for analysis of EST sequences. Nucleic Acids Res 28: 3657–3665

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 34: D332–D334

Masoudi-Nejad A, Tonomura K, Kawashima S, Itoh M, Kanehisa M, Endo TR (2004) The kingdom of Plantae EST Indices: a resource for plant genomics community. Genome Informatics 2004. PP-102. The 15th International Conference on Genome Informatics, December 16–18, 2004, Pacifico Yokohama, Japan

Masoudi-Nejad A, Tonomura K, Kawashima S, Itoh M, Kanehisa M, Endo TR (2005) GrainGenes 2.0. An improved resource for the small-grains community. Plant Physiol 138: 643–651

Schneeberger K, Malde K, Coward E, Jonassen I (2005) RepeatMasker Open-3.0. Institute for Systems Biology. http://www.repeatmasker.org (January 10, 2007)

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, et al. (1996) A gene map of the human genome. Science 274: 540–546

Smit A, Hubley FA, Green P (2004) RepeatMasker Open-3.0. Institute for Systems Biology. http://www.repeatmasker.org (January 10, 2007)

Ware DH, Jaiswal P, Ni J, Yip IV, Fan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, et al. (2002) Gramene, a tool for grass genomics. Plant Physiol 130: 1606–1613

Wheeler DL, Smith-White B, Chevretin V, Resenchuk S, Dombrowski SM, Pechous SW, Tatsuova T, Ostell J (2005) Plant genome resources at the national center for biotechnology information. Plant Physiol 138: 1280–1288

Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant Physiol 138: 27–37

Copyright © 2007 American Society of Plant Biologists. All rights reserved.