

Implementing Support for Multiple Species in XMLPipeDB's GenMAPP Builder

Don Murphy¹, John David N. Dionisio¹, Kam D. Dahlquist²

¹Department of Electrical Engineering and Computer Science, ²Department of Biology Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045 USA



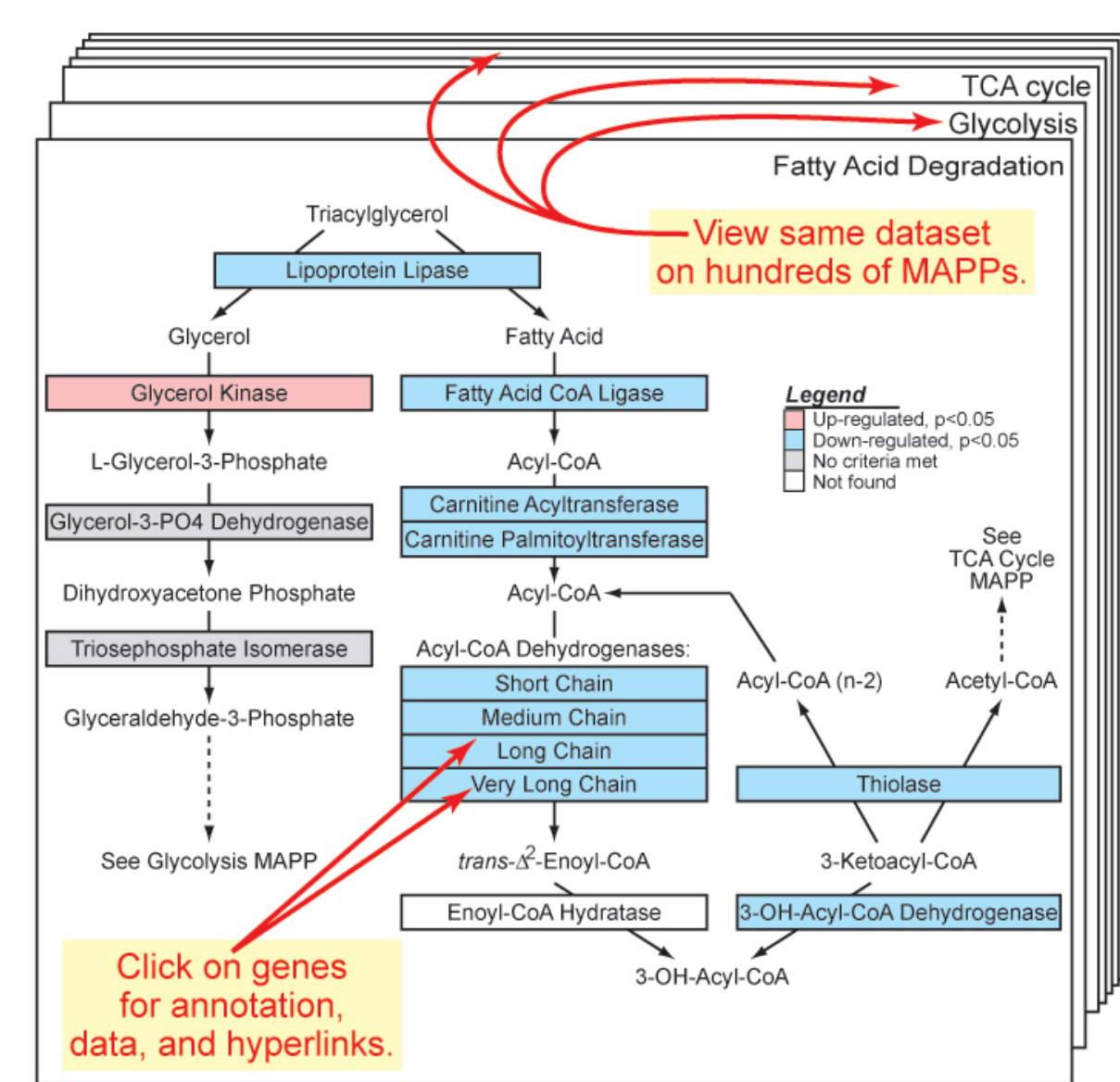
Abstract

GenMAPP Builder is a Java program used for the creation of species-specific gene databases for use with GenMAPP, a bioinformatics program for viewing and analyzing DNA microarray data on biological pathways. However, organisms with larger and more fully annotated genomes, such as *Saccharomyces cerevisiae*, may exceed GenMAPP's memory capacity, causing the program to crash. To overcome this obstacle, GenMAPP Builder has been extended. In previous builds, GenMAPP Builder had required databases to be built using all Gene Ontology (GO) terms. To allow species with larger genomes to be used with GenMAPP, exports can be performed with single subtypes of the GO annotation terms, biological process-only, cellular component-only, or molecular function-only, creating smaller gene databases. Additionally, previous versions of GenMAPP Builder held Gene Ontology association (GOA) files in memory during database export, slowing the export process and limiting the potential size of databases further. We have moved the import of the GOA linking file to the step where source data is brought into PostgreSQL, instead of the export step. This change allows the PostgreSQL database itself to be used for applications beyond GenMAPP alone. Finally, we have altered GenMAPP Builder so that it can handle importing and exporting data from multiple species at once. This functionality allows for cross-species comparison of microarray data in GenMAPP. These upgrades to GenMAPP Builder have been used in the creation of databases used in the analysis of microarray data for species such as *Saccharomyces cerevisiae* and *Staphylococcus aureus*.

Introduction

GenMAPP is a program used to view and analyze DNA microarray data

- Graphical tools within GenMAPP are used to draw biological pathways and groupings of genes with similar functions as MAPPs (.mapp).
- Expression Dataset files (.gex) store experimental data and criteria to color MAPPs.
- An underlying Gene Database stores gene IDs, annotation, and links to web-based gene and protein databases.
- The program was written in Visual Basic and reads Gene Databases in a Microsoft Access-compatible .gdb format.
- Finally, the accessory program MAPPFinder finds Gene Ontology (GO) terms over-represented in an Expression Dataset and ranks them by p-value.
- However, GenMAPP is limited because there are Gene Databases for only a few species.
- Also, the Gene Databases need to be regularly updated to include late-breaking gene annotations.



Dahlquist et al. (2002)

Figure 1. GenMAPP can be used to create MAPPs of groups of genes and then be colored to show variations in the expression of the genes.

GenMAPP Builder is a program used to create gene databases for GenMAPP

- GenMAPP Builder is part of XMLPipeDB, a suite of Java-based tools used to create databases from XML-formatted text.
- XMLPipeDB was first developed by Dr. John David Dionisio and Dr. Kam Dahlquist, and has been in development with several students since 2006.
- Maintained source files from the European Molecular Biology Laboratory (EMBL) Integ8 web portal and the Gene Ontology (GO) Consortium's website are parsed by GenMAPP Builder and inserted into a PostgreSQL relational database.
- After making associations within the database, GenMAPP Builder exports the PostgreSQL database to a GenMAPP-compatible Gene Database.
- For full functionality, each species must have a profile created in GenMAPP Builder's source code.
- However, GenMAPP Builder has several problems, such as an inefficient export method and creating databases that are too large for GenMAPP to properly use.
- To resolve these issues, changes had to be made to how one of the required source files are used and how exports are performed.

Methods

GOA source files are imported to PostgreSQL database before export

- Three types of source files are used by GenMAPP Builder to create gene databases: UniProt XML (.xml), OBO-XML (.obo.xml), and Gene Ontology Annotations (.goa).
- In previous versions of GenMAPP Builder, only the XML-formatted files were imported into the PostgreSQL database, while the tab-delimited GOA file was held in memory during export.
- An additional class was created to import GOA files into the PostgreSQL database, and the export method was altered to use the GOA data in its new location.
- Unlike the other import classes that utilized XML-to-PostgreSQL import engines that were part of XMLPipeDB, the new GOA import class had to have its own engine created to handle the tab delimited format.
- Additionally, by having GOA data within the PostgreSQL database, repeated exports can be performed in less time.

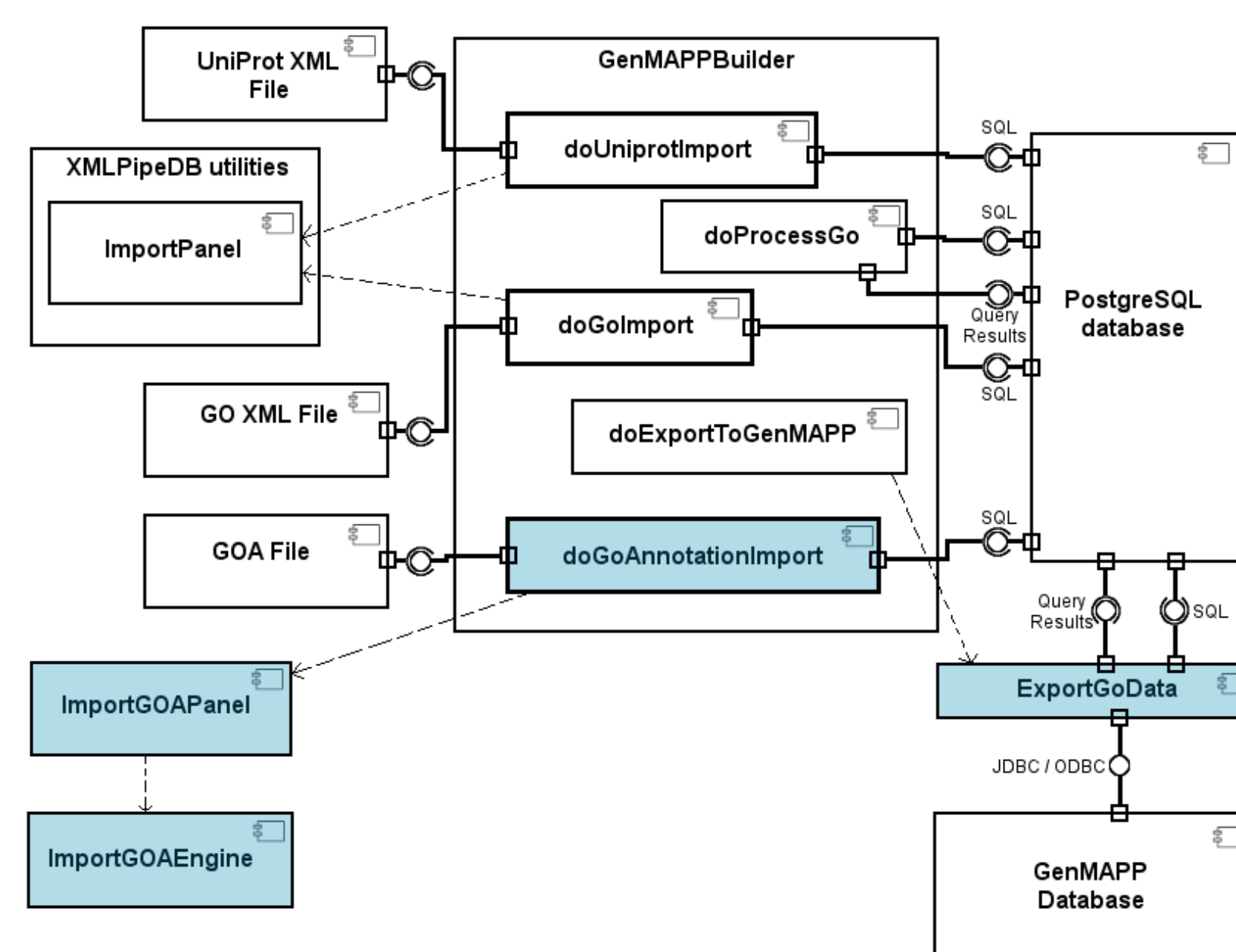


Figure 2: Components added to or edited in GenMAPP Builder for the use of GOA data within the PostgreSQL database are highlighted in blue in this component diagram.

Additional changes were needed to produce usable gene databases for well-documented species

- To create smaller but functional gene databases that could be read by GenMAPP, several methods were attempted.
- Early attempts used GO Slim variations of source files, limiting the number of GO terms added to the gene database.
- To keep all terms while producing smaller databases, exporting was modified to allow for creation of separate partial gene databases dedicated to one of the three "aspects" of GO terms.

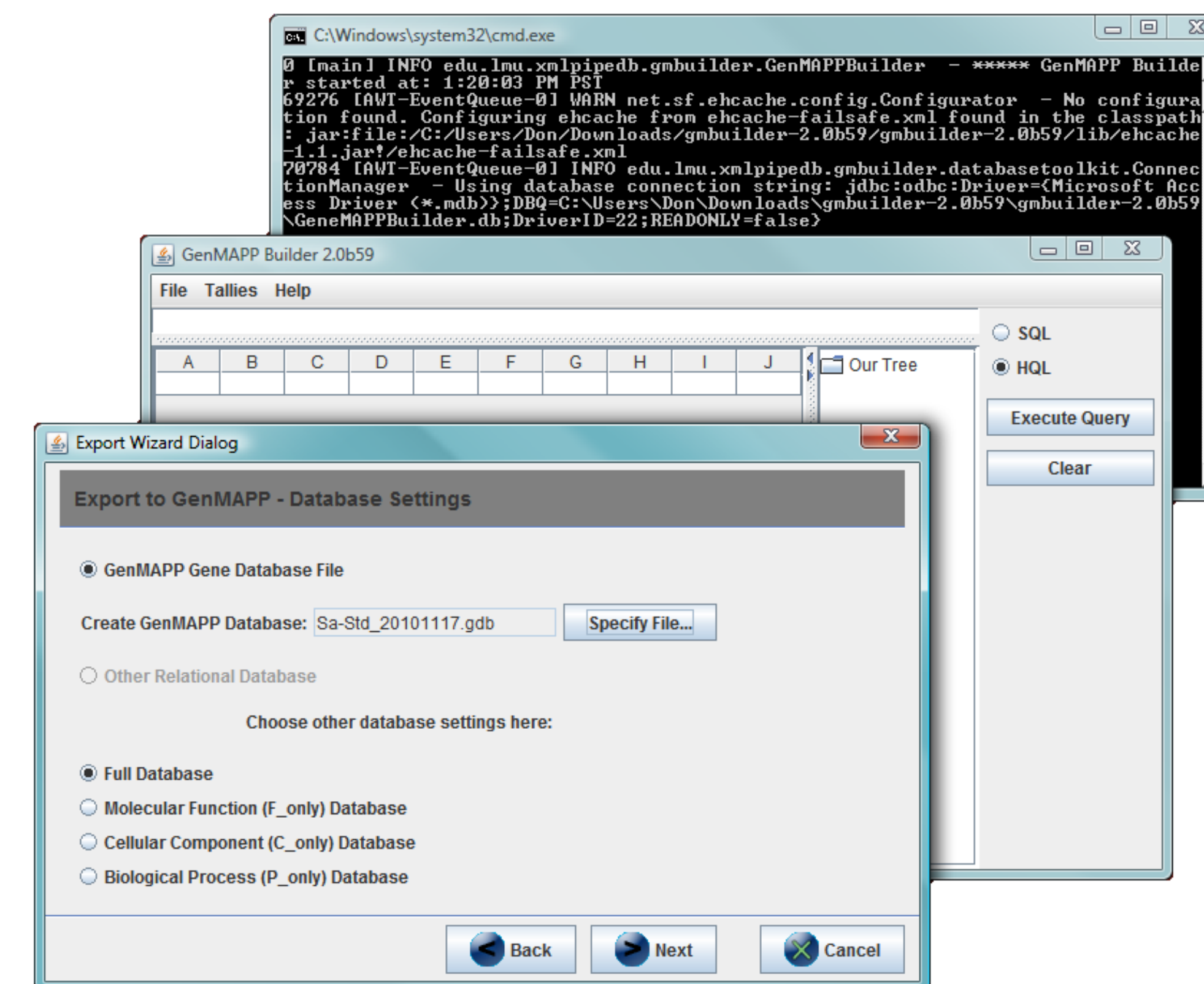


Figure 3: GenMAPP Builder's export interface lets the user choose to export either a full or single-aspect gene database.

Species profile selection and multi-species export support can be achieved by using taxon IDs

- Further examination of source code showed that all entries for all genes were being exported into the gene database, regardless of the species being exported.
- By modifying the export process to only export genes for the exported species, GenMAPP Builder could produce full gene databases that may be small enough for use with GenMAPP.
- Additionally, data for multiple species may be imported into one PostgreSQL database and then exported into different species-specific gene databases or a single multiple-species gene database.
- To avoid confusion when a species is known by two names (e.g., "*Saccharomyces cerevisiae*" and "baker's yeast"), species profiles must be changed to be identified by the unique taxonomy ID.

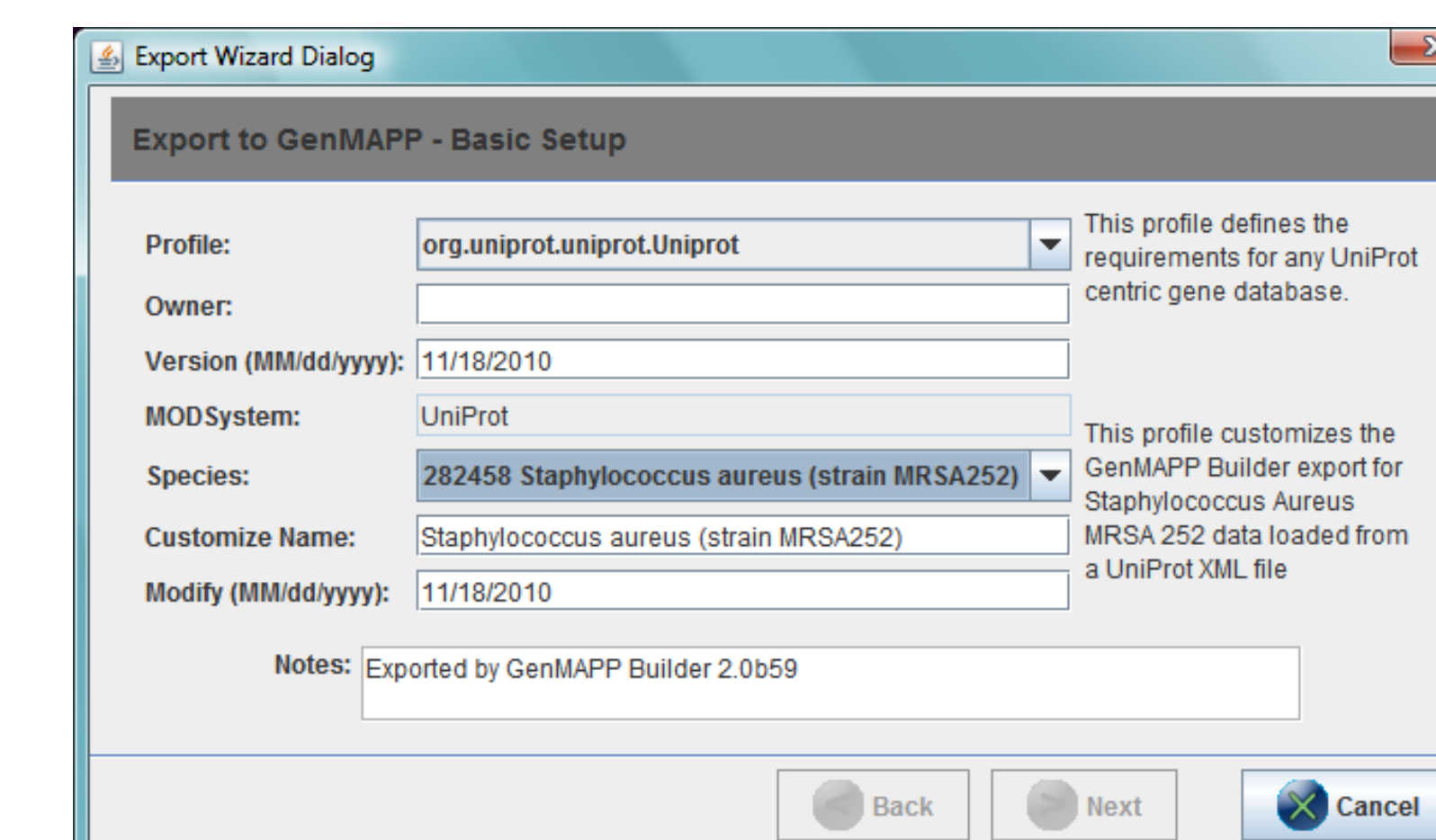


Figure 4: When selecting a species to export to a gene database, the species is identified by its taxonomy ID alongside the species name.

Results

- Improvements to GenMAPP Builder that allowed the import of GOA source files to the PostgreSQL database were effectively implemented.
- GO Slim use was experimented with, but has been rejected for the time being as the resulting databases had too many GO terms removed to be of use.
- Single aspect databases exports were performed by altering the GOA source file to include only one GO term aspect, but was added as an option during export, alongside GOA importing.
- Gene databases created by the improved GenMAPP Builder have been used in studies submitted to the Loyola Marymount Undergraduate Research Symposium 2010 and SCUR 2010.
- Multiple species import-export is still in development; recent changes which identify species by the unique taxonomy ID number as opposed to the varying species name will make such multiple-species and species-specific gene databases possible.

Discussion

Challenges Encountered

- While developing the GOA import engine, a new format for GOA files was released; to support users of both the old and new formats, the engine was developed to identify the format being used and use different insert PostgreSQL statements for each format.
- Additional troubleshooting was necessary throughout the development process; for example, an export connection error originally thought to be attributed to Windows 7 was researched and found to be caused by the inability of the program to use JDBC-ODBC database connection drivers from a 64-bit Java environment and required a fix after a campus-wide upgrade to 64-bit Windows 7.

Future Work

- As all data needed for a gene database is now present in the PostgreSQL database now that GOA files are imported, independence from GenMAPP may be achieved to avoid errors encountered with large databases due to inefficient memory use.
- A new interface could be developed to connect to the PostgreSQL database, retargeting GenMAPP's algorithms from the GenMAPP gene database to the PostgreSQL database.
- The new interface could also possibly connect to the PostgreSQL database remotely, allowing users to share the database without exchanging large files.

Acknowledgments

We would like to thank the BIOL/CMSI/HNRS 398 Biological Databases Fall 2009 Class, especially Team Elite Three (Bernadette Pak, Kenny Rodriguez). We also thank the Dahlquist lab for their extensive testing of GenMAPP Builder, as well as past GenMAPP Builder developers Derek Smith and Jeffery Nicholas.

Literature Cited

- Salomonis, N., Hanspers, K., Zambon, A.C., Vranizan, K., Lawlor, S.C., Dahlquist, K.D., Doniger, S.W., Stuart, J.M., Conklin, B.R., Pico, A.R. (2007) GenMAPP 2: New Features and Resources for Pathway Analysis. *BMC Bioinformatics* 8:217.
- Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C., & Conklin, B.R. (2003) MAPPFinder: Using Gene Ontology and GenMAPP to Create a Global Gene-Expression Profile from Microarray Data. *Genome Biology* 4:R7.
- Alphonso A, Villaflores C, Smith D, Dahlquist KD, Dionisio JDN. "Extending XMLPipeDB to create gene databases for plants and microorganisms for the analysis of DNA microarray data." First RECOMB Satellite Conference on Bioinformatics Education, March 2009.

Data Sources

- UniProt XML Proteome Sets and GOA files from the Integ8 resource <<http://www.ebi.ac.uk/integ8/>>
- Gene Ontology OBO XML from the Gene Ontology Project <<http://www.geneontology.org/GODownloadsontology.shtml>>
- Yeast-specific data from Saccharomyces Genome Database <<http://www.yeastgenome.org/>>
- Saccharomyces cerevisiae* DNA microarray data from Dahlquist lab
- map2slim from <<http://search.cpan.org/~cmungall/go-perl-0.10/scripts/map2slim>>