Updating XMLPipeDB to perform new *V. cholerae* export, then performing XMLPipeDB website updates

Nicole A. Anguiano

BIOLT 367: Biological Databases

December 21, 2016

**Abstract**

 The primary goal of this semester’s course was to complete an export of the *Vibrio cholerae* gene database, which had not been updated since 2010. In doing this, it was discovered that GMBuilder could no longer process the Gene Ontology’s OBO-XML files due to the addition of new terms and needed editing in order to be able to process them correctly. Following the edits, an export of *V. cholerae* was successfully completed and vetted for quality. After the export was complete, the XMLPipeDB website was converted to Jekyll and moved to a new github.io domain. SchemaSpy was also run to show the new structure of the GMBuilder database following all edits. Documentation was also written to allow for future edits to XMLPipeDB and SchemaSpy.

**Introduction**

*V. cholerae* is a gram-negative bacteria that causes the disease cholera, a potentially fatal disease that causes extreme diarrhea and vomiting (Finklestein, 1996). The *V. cholerae* gene database referenced by GMBuilder needed an update, as it had not been updated since 2010. However, GMBuilder was unable to export a new version of the database due to a change in the document type definition (DTD) of the Gene Ontology’s OBO-XML file. The updates to GMBuilder would be performed by using XMLPipeDB’s xsd2db and GODB Post Processor tools. XMLPipeDB is a series of tools that allow for the creation and editing of databases from XML schemas (Dahlquist & Dionisio, 2016). In addition to the GMBuilder OBO-XML processing error, the XMLPipeDB website was also out of service, likely due to a server configuration change. This made it an opportune time to upgrade the website to Jekyll, as well as update all links and materials to modern versions.

**Materials and Methods**

***XMLPipeDB Updates***

First, GMBuilder needed to be updated to allow for proper processing of the new OBO-XML files from the Gene Ontology. To begin, I ran GMBuilder and ran through the existing OBO-XML file to see what elements caused an error. Only a single term resulted in an error. I took an existing dtd file that existed on the XMLPipeDB Github (XMLPipeDB, 2016) and added the new element and subelements. Then, I ran xsd2db on this newly edited dtd file. I used GODB Post Processor on the HBM and SQL files specified by the program (To.hbm.xml and schema.sql) to allow them to be utilized by GMBuilder. Following this, I built a new godb.jar using ant from the resulting files created from xsd2db and processed by GODB Post Processor. I replaced the godb.jar in GMBuilder with the new jar file, as well as replacing the existing godb.sql file with the file processed prior by GODB Post Processor. I replaced the segments in gmbuilder.sql corresponding to the contents of godb.sql with the new contents of the new file.

***V. cholerae Export***

Following the above changes, the *V. cholerae* export could perform as is typical. I downloaded the UniProt XML file for *V. cholerae* serotype O1, version UniProt Release 2016\_09. Then, I downloaded the GOA file, version 10/4/16, 7:16:00 AM. Finally, I downloaded the Gene Ontology OBO-XML file, version 10/09/16, 3:41 pm. All files were downloaded 10/9/16. The UniProt XML and GO OBO-XML files were extracted using 7-Zip.

 Next, the GenMAPP Builder tables were created using PostgreSQL. The gmbuilder.sql file included in the GMBuilder download was run to create all necessary tables. The GO OBO-XML, UniProt XML, and GOA files were all uploaded using GMBuilder, then exported into a GenMAPP gene database (.gdb). The resulting database could then be vetted for quality.

***Website Changes***

The site was built initially using Maven so that its initial appearance could be copied in Jekyll. The structure of the Maven site was mirrored in the Jekyll site. Following the completion of the site and updating of all links and outdated content, SchemaSpy was run. A SchemaSpy tutorial can be found on the XMLPipeDB github (Anguiano, 2016).

**Results**

 Due to the website changes and XMLPipeDB updates not having any tangible results, results will only be given for the *V. cholerae* export. All figures are displayed at the end.

 According to TallyEngine (Fig. 1), there were 3,380 OrderedLocusName IDs in both the database and XML. This number was not matched by those returned by XMLPipeDB match utlity nor the Original Row Counts Table in the GDB itself (Table 1). The XMLPipeDB match utility used the following command to generate its results: java –jar xmlpipedb-match-1.1.1.jar “VC\_A?[0-9][0-9][0-9][0-9](.1|)” < uniprot-organism%3A243277.xml. This returned 3,832, not 3,830 as may have been expected. A SQL query performed directly within the database (select count(\*) from genenametype where type = ‘ordered locus’ and value ~ ‘VC\_A?[0-9][0-9][0-9][0-9]’;) returned 3,830 Ordered Locus Names as well. The reason for the discrepancy was due to formatting in PostgreSQL. Two cells in the database contained two geneIDs each: VC\_1738/VC\_1739 and VC\_A0015/VC\_0016. When these two cells are separated, the above query returned the expected number: 3,832.

 GenMAPP and MAPPFinder filtering has not yet been performed due to the inability of GenMAPP to run on Windows 10. No changes were made to the GMBuilder code outside of what was required to allow for the export.

**Discussion**

 GMBuilder was not capable of performing a *V. cholerae* export at the beginning of the semester due to formatting changes. As a result, the GenMAPP Builder process was quite difficult at first. However, once changes were made that allowed it to process successfully, the process was quite smooth and occurred without difficulty. The export seemed of good quality and to have no errors.

 The changes to the XMLPipeDB site are visible on the new github.io domain at http://lmu-bioinformatics.github.io/xmlpipedb/. The new SchemaSpy documentation is also visible there.

**Conclusions**

 To further this project, I would like to make additional changes to the site so as to update the content further. Additionally, I would like to perform more exports on more species that have not been updated in quite some time. I would also like to try to find a way to allow GenMAPP to run on Windows 10, as it crashes if an expression dataset is created.

**Acknowledgements**

 I would like to acknowledge Dr. Dahlquist and Dr. Dionisio for giving me the opportunity to continue this project this semester and for guiding me as I completed the tasks I was given.

**References**

Anguiano, N. A. (2016, November 28). Using SchemaSpy. Retrieved December 19, 2016, from https://github.com/lmu-bioinformatics/xmlpipedb/wiki/Using-SchemaSpy

Dahlquist, K. D., & Dionisio, J. N. (2016, November 07). XMLPipeDB - About. Retrieved December 19, 2016, from http://lmu-bioinformatics.github.io/xmlpipedb/about.html

Finkelstein, R. A. (1996). Cholera, Vibrio cholerae O1 and O139, and Other Pathogenic Vibrios. Retrieved December 19, 2016, from https://www.ncbi.nlm.nih.gov/books/NBK8407/

**Figures**

**Figure 1: TallyEngine Results**



Fig. 1: TallyEngine results. 3,830 Ordered Locus names were found in both the XML and the Database. The number of GO terms was also the same, at 45,553. ‘

**Table 1: OrderedLocusNames IDs**

|  |  |
| --- | --- |
| **Type** | **Number of IDs** |
| XMLPipeDB match | 3832 |
| TallyEngine (UniProtXML) | 3830 |
| TallyEngine (Database) | 3830 |
| OriginalRowCounts Table in GDB | 7664\* |

Table 1: Depicts the number of OrderedLocusName IDs as obtained by XMLPipeDB match utility, TallyEngine (for both UniProtXML and the Database – see Fig. 1), and the OriginalRowCounts Table in the GDB. Note that the OriginalRowCounts Table in the GDB contains duplicates for every term. There are only 3,832 unique terms in the OriginalRowCounts table. XMLPipeDB match results were generated using the command: java –jar xmlpipedb-match-1.1.1.jar “VC\_A?[0-9][0-9][0-9][0-9](.1|)” < uniprot-organism%3A243277.xml