Individual Assessment and Reflection

Lena Hunt

BIOL 367: Biological Databases

December 8, 2013

Statement of Work

For team *Off the Leish,* I operated as the Quality Assurance member. I ran multiple import-export cycles with our team Coder, Gabe. Together, Gabe and I maintained a regularized system of file management. Our system was to store all downloads in the Keck Downloads folder. Data was named *"Leishmania_(name of file)_(MMDDYYYY)_Lena_Gabe.extension.* This came in handy after many export cycles to keep track of which files were the most recent versions. It was also possible to trace the files back to their release dates when writing the group paper. All downloads were also uploaded to our wikipage for easy access from other computers and to make them available for our GenMAPP users, Viktoria and Kevin. As the QA person, I made myself knowledgable about the system IDs. Unlike other species, *Leishmania major* did not list IDs under OrderedLocusNames, instead they were categorized as ORFs (Open Reading Frames.)

After a successful export cycle, I performed part of the Gene Database Testing Report. Using TallyEngine, I compared what was counted in the Database to what was counted in the XML file. It took several exports to get them to match, and to have GO terms show up. Gabe and I also customized the TallyEngine setup for our species to have and additional column to recoginize ORFs since we did not have OrderedLocusNames. I used XMLPipeDB Match to characterize a regular expression pattern. This again took several tries as our ORFs were varied and any individual ORF might contain periods, spaces, underscores, or a combination thereof. With the code **[Ll][Mm]j[Ff][_ .]##[_ .]###** I managed to catch 8353 of the 8355 of the ORFs. Opening the query up any more yielded a much higher number of results that included non-relevent data. Using Direct SQL queries, the command **select count(*) from genenametype where type='ORF' and value ~ 'L[Mm][Jj]F(.|)[0-9][0-9].[0-9][0-9][0-9][0-9]'** found 8350 of the 8355 ORFs. The final 5 stragglers were picked up by the command **select value from genenametype where type='ORF' and not value ~ 'L[Mm][Jj]F([\. _]|)[0-9][0-9][\. _][0-9][0-9][0-9][0-9]'.** I opened the Gene Database in Microsoft Access to perform an Original Row Counts comparision. There were 16662 Original Rows, which is roughly twice the number of ORFs. This is due to a change in the code that normalized the gene IDs; therefore many of the gene ID showed up twice, once in their original format and once in their normalized format.

Gabe and I worked to create a database schema base on the Vibrio cholera schema, which ended up being the same scheme as the Vibrio cholera one. We also wrote up the readme file for the project.

Assessment of Project

I think, that given our amerature status as compueter scientists and our brand new knowledge of bioinformatics, we were able to work very successfully on the project. We did our best to stay ahead so that when we ran into stumbling blocks, we had time to ask for help. I think our team shared work well and supported each other when we ran into problems.

In what we thought was our final version of the database, the export didn't end up doing what we thought it was going to do. When we went back to look at the code, there was a minor error (a * in lieu of a %.) We ended up having to re-export with the code adjustment at the last minute. In our final version of the database, the code worked and the database was able to find almost all of the IDs that were present in the XML file.

If I could do it again, I would make fewer exports. As a group, we exported so many versions of our database that we ended up losing a lot of class time waiting. It was definitely a learning process, but now that I am more aware of what should be present, I think I could save unnecessary exports.

I think the work is high quality, we managed to include almost all of the genes in our final export. I think the amount of time we put into generating this database really paid off in its completeness. The last few gene IDs were so varied it would be difficult to capture them with computer code.

I don't think our team was the most organized, and our wiki page was very hap-hazard for a while. I went back in to organize it a little better, regulate what was bolded and italicized and so on. Meeting up to work on the project was also very disorganized, but that was simply due to the fact that we all have busy schedules. I also think our PowerPoint could have been better organized if we had more time. We were working on it last minute and only ran through our slides once.

Yes, I believe *Off the Leish* did achieve all of its objectives for this project. We were able to complete at least the base levels of what was required, and I think all my teammembers were able to meet the milestones for their roles in the group.

Reflection on the Process

With my head I reviewed the way genes work and the principles of biological pathways. I learned the process of making a new gene database very well, and I think I could probably easily re-do my portion of this project for another species without any confusion or problems. The whole process of managing and storing data to use later was totally new to me, but by the end I finally felt like I understood the process.

With my heart I learned that leaving group work till the middle of finals week does not work. Everytime we met up to work on the presentation, at least one person was exhausted from taking a final or hadn't eat because of studying. Given that we were working on out database up until the last week of classes, I am not sure how we could have avoided this, but perhaps getting more done over the weekend would have helped. Using Google Docs so that group members could edit the paper and PowerPoint individually and from their respective homes was really helpful, especially since most of our group were off-campus.

With my hands I learned how to do basic coding, which was very satisfying. I was able to find gene ID is XMLPipeDB Match and SQL queries. I also got into the habit of recording what I was doing in an online journal as often as I could. This was a hard habit to adapt because it is just so much quicker to go without recording things, but having a log of what I did was really helpful later.

It may be a small thing, but I have started adding dates and more specific title to all the documents on my own computer. I have found that saving multiple copies of documents with different dates makes it easier for me to go back and make edits, especially when I am working on semester-long projects. I also now add my last name--which I never did on my own computer--but when I send assignments to professors I don't have to worry about them mixing up my paper. Basically, my personal computer habits are a lot more organized and I understand how my computer works a little better.