Generating a New Gene Database for Leishmania major using GenMAPP and XML pipedb

Gabriel Leis, Viktoria Kuehn, Lena Hunt, Kevin McGee

BIOL 367: Biological Databases

December 8th, 2013

Introduction:

Leishmania major is a tropical species of protozoa that causes a range of human diseases known as leishmaniases. Leishmaniases affect 2 million people in 88 countries annually. From the paper and accompanying microarray data by Ivans et. al, it is know that the Friedlin strain of Leishmania major contains 36 chromosomes with a 32.8-megabase haploid genome. The genome contains an estimated 911 RNA genes, 39 pseudogenes, and 8272 protein-coding genes (of which up to 36% are believed to have a specific function.) This information was derived from a DNA microarray experiment. The microarray experiment was designed to reveal differences in gene expression between the promastigote and amastigote life cycle stages of Leishmania major. At the start of this project, the data collected in the microarray experiment was unable to be analyzed by GenMapp/MAPPFinder due to the fact that there was no Gene Database for Leishmania major. This project used XMLpipedb, an open source program for building relational databases from an XML schema, and GenMAPP Builder, a program for creating GenMAPP database files, to generate a new database. The newly created database allows the microarray data from Ivans et. al to be analyzed using MAPPFinder, a tool that creates gene-expression profiles using annotations from the Gene Ontology program. Using MAPPFinder, GO terms with over-represented gene-expression changes may be found and displayed in a graphical, searchable, and annotatable file.

Methods:

Data Source Files

The Uniprot XML proteome set was downloaded from the Uniprot complete proteomes page for Leishmania major. We used the version that was last updated on October 16th 2013. The GOA (GO association) file was downloaded from the Uniprot-GOA downloads page. Our version was downloaded on November 14th 2013. The GO file was downloaded from the Ontology Downloads page (we used the beta version of the page). The version was from November 4th 2013, 2:03:38 AM in the obo-xml.gz format.

Generating Gene Database

In PostgresSQL, a new database was created called Leishmania_05112013_Lena_Gabe. GenMAPP Builder tables were then created in PostgresSQL. GenMAPP builder was downloaded from Source Forge (version: GenMAPP Builder 2.0b71.) The Leishmania database that was created in PostgresSQL was then configured to GenMAPP Builder and we imported our Uniprot XML and GOA files and our GOA file. Once all the files were imported, we exported a GenMAPP Gene Database for Leishmania major was saved as Leishmania_05112013_Lena_Gabe.gdb.

Inspecting Database

To make sure our Gene Database export had worked, we ran Tally Engine to make sure our XML count matched our Database count. We also used XMLpipedb Match and SQL query to make sure we could match our Gene IDs.

Prepare microarray data (organize, normalize, perform statistical analysis)

In order to analyze the microarray data the data files from the experiment fisrt had to be prepared for GenMAPP. The data that corresponds with the paper was found in ArrayExpress and downloaded in the .txt format. The files that were used were the Raw Data File and the Sample and Data Relationship File (SDRF). Because this paper compared the promastigote and amastigote stages in two different species the data had to be reorganized. The SDRF file was opened in excel and

rearranged so that the chips were grouped for each species. Then the Raw Data file was opened in excel and the IDs on the chips were matched to the SDRF file. Leishmania major raw data was prepared separately from infantum on two individual excel spreadsheets from this point on. L. major had 6 chips while L. infantum had 8.

The raw data could now be organized and filtered down to only the relevant information for statistical analysis and GenMAPP. The raw data that corresponded to the name of the ID for both species was found and labeled with the last two digits of the ID in it. Only the name (now with the number in the header of the columns) and the expression ratio were kept for each chip. Some of the dyes were swapped, so these were inverted to match the rest of the values. The data was then normalized. New columns for the standard deviation, and average were calculated from each sample's microarray ratio to scale and center the data. The average log full change was then calculated and statistical tests were run. A t-test and p-value was calculated for each of the samples, which were analyzed for relevance. The data was prepared for GenMAPP by adding a column called System Code and the decimal places were formatted to be compatible with GenMAPP. The data was now formatted and ready for import into GenMAPP.

Running GenMAPP using the Gene Database

Microarray data (import using Expression Dataset Manager)

GenMAPP was launched. Under the Data menu, the Leishmania major database was selected. Again under the Data menu, Expression Dataset Manager was opened. The tab-delimited text file formatted for GenMAPP was selected for the file. Expression Dataset Manager converting the data to a .gex file. There were 1820 lines of data that could not be converted. These lines of data were exported in an .EX.txt file (exceptions file.)

Run MAPPFinder analysis

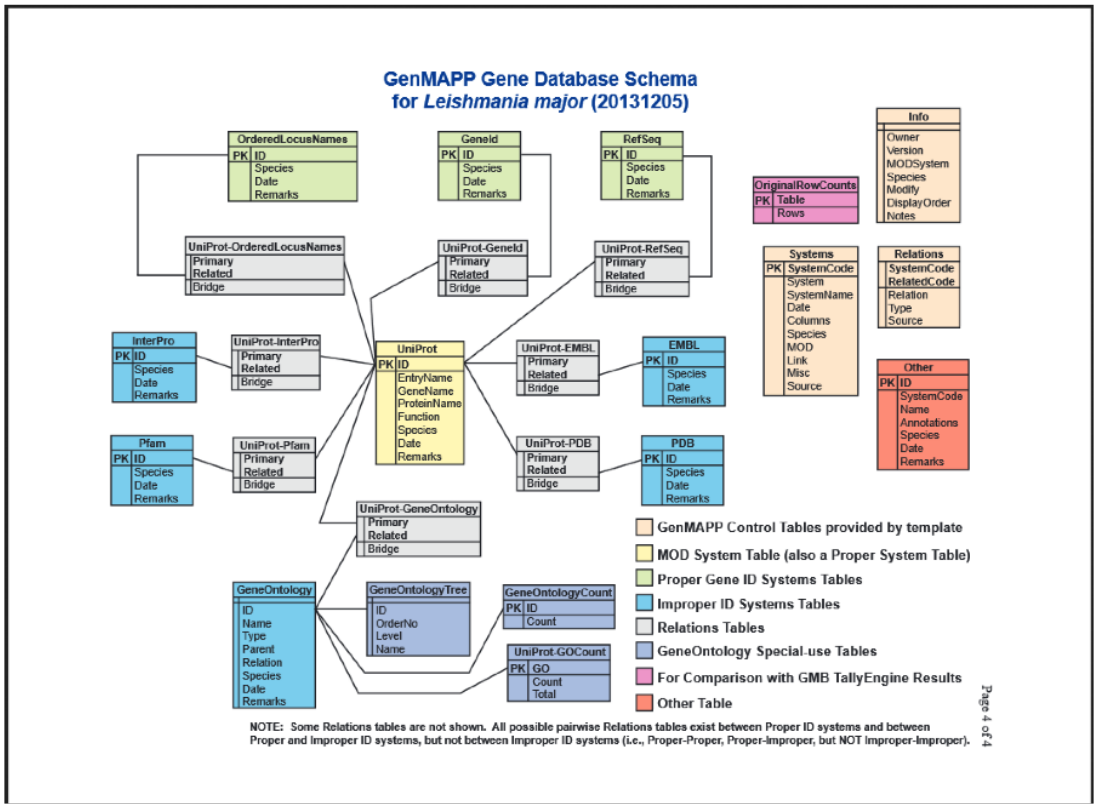
After this data was uploaded, we prepared the data for MAPPFinder by filling in the color set area of the expression dataset manager. The colors were assigned to the two main criterion: Increased relative to control had a Log FC> 0.25 and P-Value <0.05 these were colored blue. Decreased relative to control had a Log FC< -0.25 and P-Value <0.05 these were colored purple. Saved this color set under the title "Amastigote vs. PRoma".

Next, we went to tools on the GenMAPP homepage and ran MAPPFinder off of there. This brought up a pop-up screen where we plugged in our data. We chose "Amastigote vs. PRoma" as our color set to be used, clicked the buttons "Gene Ontology" and "Calculate p-values". Finally, we clicked run MAPFinder. Saved the file of the gene ontology table as LMajorGOMap. MAPPFinder ran for about an hour and a half. When it was done, we had our complete gene ontology table which we were able to use to draw a MAPP pathway.

The significant GO term "Aromatic Compound Catabolic Process" was used to create a comparative pathway map as a representative map for other terms found in MAPPFinder. The original term was opened so the map of genes could be viewed. Genes that were not found were removed from the map. Significant genes were that researched on Uniprot to figure out their purpose, and they were then oriented on the map to make a clear representation of the differences in the genes.

Results:

Gene Database Schema



Gene Database Testing Report on final version of Gene Database

In GenMAPP Builder, LeishmaniaGDB_Lena_Gabe_20131203.gdb was created. TallyEngine matched 8041 UniProt IDs, 0 Ordered Locus Names, 8355 ORFs, 8317 RefSeq IDs and Gene ID Ids, and 40065 GO Terms. XMLPipeDB Match was used to validate XML results from the TallyEngine. XMLPipeDB Match managed to capture 8353 ORFs, leaving 2 ORFs that could not be captured by coding. SQL Querie was used to validate the PostgreSQL Database Results from the TallyEngine. Using 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]';** 8350 ORFs were captured. The five straggling ORFs were captured 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]';.** The OFRs that were not captured by the initial command were: LMAJ006828, L1063.01, L3640.11, Lmj 1130, L374.02.

Report on quantity and identity of gene IDs that did not make it into the database

1820 genes were recorded in the GenMAPP exceptions file. Of these, 1753 IDs were not present in the XML source at all. These genes followed the ID pattern LmjF01.###[1-9]. The exceptions to this rule included any ID ending in "0" as well as any ID that had a lower number than LmjF01.0160 or a number greater than LmjF01.1983. All IDs present in the XML were found in Postgres. However, some IDs found in Postgres were not exported to GenMapp. Some of these IDs are not present in the XML while other IDs not present in XML follow the form LmjF01.[0160-1983]The IDs ending in zero are found in XML for example IDs Lmjf01.0160 and LmjF01.1970 are found but IDs LmjF01.016[1-9] and LmjF01.197[1-9] would not be found. There are also five extreme outliers that were not exported to GenMAPP. These IDs are LMAJ006828, L1063.01, L3640.11, Lmj 11430, and L374.02.

Report on what changes need to be made to the GenMAPP Builder code in order to to accommodate the second and third type of missing gene IDs

Initial changes were made in the database code to instruct Postgres to obtain IDs from 'ORF' rather than ordered locus names. Several changes were made to the custom species profile to accommodate a variety of IDs found in the microarray data. The code was normalized data so that each ID follows two forms: LMJF.##.#### and LMJF_##_####. Some IDs are still not being exported that follow a distinct pattern of LmjF01.####. Further work needs to be done to adjust the code to accommodate these as well as some distinct outliers.

Report results of the DNA microarray analysis

The results from the DNA microarray analysis showed to have significant change in gene expression. The first time the microarray statistics were analyzed, the excel spreadsheet used did not have the repeat samples removed and did not have the L. infantum results properly filtered. This resulted in a number of statistically significant changes in gene expression that was too high for only the L. major data. Once these were properly filtered and the significant p-values were found again the results made more sense. With the p-value cut off of less than 0.05 there were a total of 514 genes found that had significant changes in gene expression out of a total of 912 *L. major* samples (with repeats of each for a total of 1824). The criteria for a significant increase was a log fold change of greater than 0.25 and a p-value of less than 0.05. This resulted in a 17.59 percent of significant changes in gene expression between the life stages of *L. major*. The significant decreases in gene expression were the ones with an average log fold change of less than -0.25 and a p-value of 0.05. The number of significant increases found were 321 and the number of significant decreases found were 169:

MAPPFinder results

Significantly Increased Gene Ontology Terms

GO Name	Z Score	PermuteP
oxidoreductase activity, acting on the CH-NH group of donors	3.003	0.009
endopeptidase activity	2.468	0.018
intrinsic to membrane	2.426	0.023
integral to membrane	2.426	0.023
autophagy	2.368	0.029

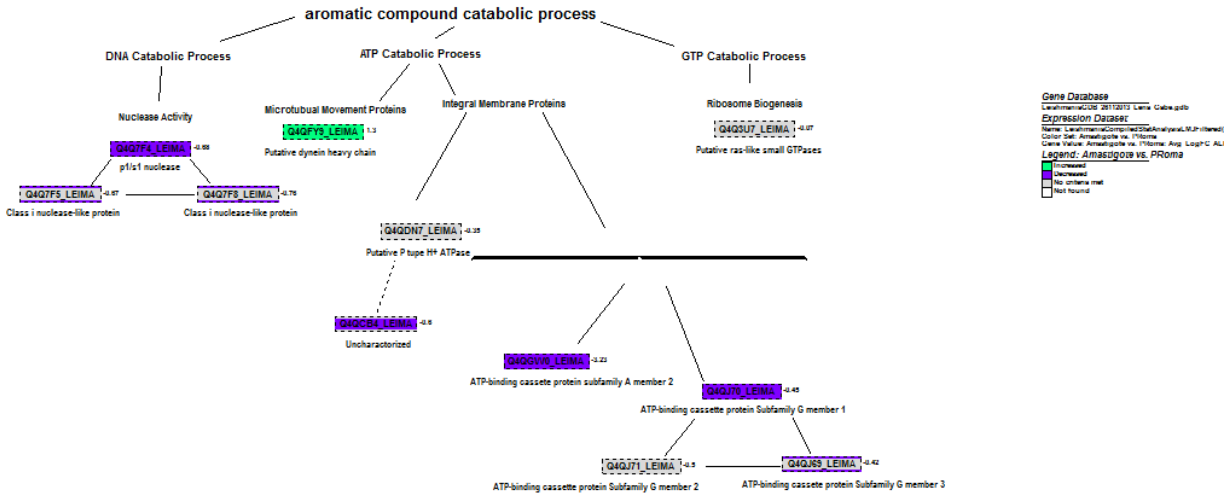
This is a table showing the Gene Ontology Terms found to have the greatest increase of fold change. This table was taken directly from the text file created by MAPPFinder. In this table, the term "Intrinsic to Membrane" is the parent term of the term "integral to membrane".

Significantly Decreased Gene Ontology Terms

GO Name	Number Changed	Z Score	PermuteP
cellular nitrogen compound catabolic process	7	3.599	0.002
aromatic compound catabolic process	7	3.599	0.002
nucleobase-containing compound catabolic process	7	3.599	0.002
organic cyclic compound catabolic process	7	3.328	0.002
heterocycle catabolic process	7	3.328	0.002
catalytic activity	46	3.153	0.002
hydrolase activity	20	2.564	0.011
organic substance catabolic process	8	2.214	0.038
ATPase activity	6	2.178	0.041

This is a table showing the Gene Ontology Terms found to have the greatest decrease of fold change. This table was taken directly from the text file created by MAPPFinder. In this table many terms are related by the supercategory of catabolic processes. On a smaller scale, "Organic Cyclic Compound Catabolic Process" is the parent term to "Aromatic Compound Catabolic Process", the term used to create the GenMAPP comparative pathway map.

GenMAPP pathway of Aromatic Compound Catabolic Process



This is a comparative pathway map of the GO term Aromatic Compound Catabolic Process. We chose this term because, not only is it significant, it is also connected to many of our other significant GO terms, including multiple catabolic process terms (DNA catabolic process, organic substance catabolic process, etc.), membrane terms, and so on. Further, this term is the offspring term of "Organic Cyclic Compound

Catabolic Process", another significant GO Term. This map shows that almost all catabolic processes involving ATP is decreased.

Discussion

GenMAPP builder process was difficult for several reasons. First, a wide variety of ID formats necessitated several changes in the custom species profile to accomodate the IDs. Secondly, many IDs missing in the XML which caused difficulties interpreting the errors when importing the data into GenMAPP. Editing of the species profile was needed for Tally Engine to correctly identify the number of genes in the profile. The gene IDs were also difficult to place in XML and identify using match and PostgreSQL. These issues all resulted in multiple exports necessary to produce the final database.

When comparing the results from the statistical analysis and the MAPPFinder in this project to the results for the microarray dataset, it was found that they were related. In the original paper comparing the differences in gene expression between different developmental life stages of L. infantum and L. major it was found that there were variations in gene expression found primarily in metabolism, cellular organization and biogenesis, and transport Gene Ontology categories. It found that 25% of differently expressed genes between life stages were involved in metabolism in both species (Rochette et al., 2008). Promastigotes had upregulated genes involving carbohydrate and glucose metabolism. A previous study mentioned in their discussion found that Leishmania species use different main sources of energy in the different life stages. During the promastigote life stage the main energy source is glucose, which shifts to fatty acids and amino acids in the amastigote phase (Rochette et al., 2008). The results from the MAPPFinder and the top Gene Ontology terms found to have changed in this project correspond to the results in the paper. The main gene ontology terms that were found in this analysis that were close in relation were ones that were under the GO category of catabolic process and metabolic function. These changes in expression when comparing the two life stages are likely to account for the changes in expression of proetins that help metabolize the main energy source. Another general difference found in gene expression between life stages of L. major were under the cellular organization Gene Ontology category. The results in changed gene expression having to do with membrane function may support this result from the paper. The cells membrane function and cellular organization changes from motile flagellated promastigotes to nonmotile amastigotes. The environment in which these two live in are very different as well; the promastigote lives in the gut of the insect host, while the amastigotes reside in the macrophage of the mammalian host (Rochette et al., 2008). These differences in environment and cellular structure are likely reasons why there were changes in expression found both in the project and in the article that have to do with membrane function and organization. The fact that this project yielded results that support the results found in the article both assures the quality of their analysis of the data, and sheds light into specific genes and functional areas that have changes in expression between life stages in *Leishmania major*. Further research may be done to compare the results of *Leishmania infantum* in the same manner to look at the extent of the similarities in expression changes between the developmental stages of the two different species.

References:

Alasdair C. Ivens, Christopher S. Peacock, Elizabeth A. Worthey, Lee Murphy, Gautam Aggarwa, Matthew Berriman, Ellen Sisk, Marie-Adele Rajandream, Ellen Adlem, Rita Aert, Atashi Anupama, Zina Apostolou, Philip Attipoe, Nathalie Bason, Christopher Bauser, Alfred Beck, Stephen M. Beverley, Gabriella Bianchettin, Katja Borzym, Gordana Bothe, Carlo V. Bruschi, Matt Collins, Eithon

Cadag, Laura Ciarloni, Christine Clayton, Richard M. R. Coulson, Ann Cronin, Angela K. Cruz, Robert M. Davies, Javier De Gaudenzi, Deborah E. Dobson, Andreas Duesterhoeft, Gholam Fazelina, Nigel Fosker, Alberto Carlos Frasch, Audrey Fraser, Monika Fuchs, Claudia Gabel, Arlette Goble, André Goffeau, David Harris, Christiane Hertz-Fowler, Helmut Hilbert, David Horn, Yiting Huang, Sven Klages, Andrew Knights, Michael Kube, Natasha Larke, Lyudmila Litvin, Angela Lord, Tin Louie, Marco Marra, David Masuy, Keith Matthews, Shulamit Michaeli, Jeremy C. Mottram, Silke Müller-Auer, Heather Munden, Siri Nelson, Halina Norbertczak, Karen Oliver, Susan O'Neil, Martin Pentony, Thomas M. Poh, Claire Price, Bénédicte Purnelle, Michael A. Quail, Ester Rabbinowitsch, Richard Reinhardt, Michael Rieger, Joel Rinta, Johan Robben, Laura Robertson, Jeronimo C. Ruiz, Simon Rutter, David Saunders, Melanie Schäfer, Jacquie Schein, David C. Schwartz, Kathy Seeger, Amber Seyler, Sarah Sharp, Heesun Shin, Dhileep Sivam, Rob Squares, Steve Squares, Valentina Tosato, Christy Vogt, Guido Volckaert, Rolf Wambutt, Tim Warren, Holger Wedler, John Woodward, Shiguo Zhou, Wolfgang Zimmermann, Deborah F. Smith, Jenefer M. Blackwell, Kenneth D. Stuart, Bart Barrel, Peter J. Myler (2005) The Genome of the Kinetoplastid Parasite, Leishmania major. Science;309(5733):436-42.

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* .

Rochette, A., Raymond F., Ubeda JM., Smith M., Messier N., Boisvert S., Rigault P., Corbeil J., Ouellette M., Papadopoulou B. Genome-wide gene expression profiling analysis of Leishmania major and Leishmania infantum developmental stages reveals substantial differences between the two species. BMC Genomics, (2008 May 29). **9**:255.

Acknowledgments

We would like to thank Dr. Dahlquist for her help and support in understanding bioinformatics and microarray data, and Dr. Dionisio for his continued support with the programing aspects of this project.