Virginia Commonwealth University

Master of Science Thesis

A Pipeline for Creation of Genome-Scale Metabolic Reconstructions

Author: Shaun Norris
Shaun William Norris

Supervisor: Dr. Paul Brooks
Dr. Paul Brooks

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Bioinformatics

in the

Paul Brooks Lab
Bioinformatics

December 13, 2016

VIRGINIA COMMONWEALTH UNIVERSITY

Abstract

Dr. Paul Brooks
Bioinformatics

Master of Science in Bioinformatics

A Pipeline for Creation of Genome-Scale Metabolic Reconstructions

by Shaun William Norris

The decreasing costs of next generation sequencing technologies and the increasing speeds at which they work have lead to an abundance of 'omic datasets. The need for tools and methods to analyze, annotate, and model these datasets to better understand biological systems is growing. Here we present a novel software pipeline to reconstruct the metabolic model of an organism *in silico* starting from its genome sequence and a novel compilation of biological databases to better serve the generation of metabolic models. We validate these methods using five *Gardnerella vaginalis* strains and compare the gene annotation results to NCBI and the FBA results to Model SEED models. We found that our gene annotations were larger and highly similar in terms of function and gene types to the gene annotations downloaded from NCBI. Further, we found that our FBA models required a minimal addition of transport reactions, sources, and escapes indicating that our draft pathway models were very complete. We also found that on average our solutions contained more reactions than the models obtained from Model SEED due to a large amount of baseline reactions and gene products found in ASGARD.

Acknowledgements

The author wishes to thank several people who were pivotal in the completion of this project and his education. First, I would like to thank my advisor and mentor Dr. Paul Brooks whose guidance and oversight were a key element in my journey. Dr. Stephen S. Fong and Dr. Maria C. Rivera for participating on my committee and whose insight and instruction were pivotal in my success. Next, Dr. Allison Johnson and Dr. Herschell Emery whose constant positive encouragement and advice kept me progressing forward. Dr. Jeff Elhai for being my toughest critic and driving me to really think critically about everything I thought I knew.

I'd also like to thank Carlisle Childress and the other Center for High Performance Computing staff for all their hard work and helpfulness throughout the years.

Last but not least, I would like to thank my family. My wife, AnaClarissa and son Harrison Norris for supporting me throughout this long and challenging journey. My parents James and Katina Norris, my brother Jesse and his wife Alix Norris who have taught me so much and encouraged me to learn more, do more and be a better person. Without all of their love and continued support, I would not have been able to do this.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Historically the hallmark of biology has been the study of the individual molecular components that make up living organisms. However, since the advance of sequencing technology and high performance computing this paradigm has shifted to a more complete approach in which a biologist considers the biological networks that make up the systems that regulate and sustain life for an organism. Continued research into genomes, gene expression and regulation continues to develop and with it so does our understanding of how each of the elements of an organism interact with one another.

Along with this systematic, holistic approach to understanding biological complexity and an increase in computational power have lead to the emergence of new methods for modeling these networks. By using mathematics one can now represent a metabolic pathway and simulate dynamic and complex biological cellular behaviors. The ability to experimentally obtain genomic data coupled with these modeling approaches has lead to a top-down approach in which the experimental data can be integrated with the models. This lends greater credibility in the models themselves and the ability to more accurately represent life *in silico*.

The ability to represent an organism *in silico* has allowed research to be conducted without the overhead costs associated with a traditional experiment. Using computers you can now predict the outcomes of gene knockouts, and gene up/down regulation. You can also identify drug targets and study complex pathways to identify methods to turn them on, off or bypass them completely. All of this together has lead to a deepening of our understanding of biology and increased the effectiveness of traditional experiments while reducing costs.

However put into the time line of biological research using *in silico* modeling is still very new and was initially cost prohibitive. It takes a lot of computational overhead to be able to perform these types of experiments. First, it requires databases that contain experimentally obtained information. Databases hosted by NCBI and other resources are freely available and often can be accessed without downloading the entire dataset. While other databases are proprietary and require you to download them before using them. Next, the sheer amount of data generated by sequencing, genome annotation, and modeling does require a lot of physical disk space. Most of the intermediate files can be compressed or removed after a functional model is produced but even this can require gigabytes of space per model. Finally, computational time, as in the actual cost of CPU usage while the modeling procedures are running can also be quite high but due to parallelization and job queuing engines like Sun GridEngine utilizing large amounts of computational time for generating *in silico* has become much easier.

The major contributions from our work here is the ability to start from the nucleotide sequence and use our pipeline in a semi-automated fashion to reconstruct the metabolic networks of a given organism. Previously this process was laborious and there were no tools to parse a genome annotation and derive its reactions based on the gene products determined during the annotation. Further, we completely redesigned the MetModel into a software tool that will execute all of the necessary steps to create a constraint-based model using flux balance analysis (FBA) all the way to generate the KGML pathway maps from a single execution of the new tool.

1.1 Reference Databases

Most *in silico* tools and projects involve using a reference database at some point. The internet is a great method for sharing information from databases and a number of biological databases already exist to share information about genes, metabolites, and their reaction pathways (Keseler et al., 2013).

However, there is no standardization among these databases and often minimal curation of the data once it has been made available. So not only is there no one source of information that houses all of the data but there is also no standardized form for the information in these databases. This is particularly true when it comes to the metabolic reactions and their metabolites. This makes it difficult to verify the data in these reference databases and it also makes it difficult to obtain a consensus of information from these databases as comparing them is often difficult.

Kyoto Encyclopedia of Genes and Genomes (KEGG), is one such database that contains genes, metabolites, reactions, and more for many different organisms(Ogata et al., 1999). However, it lacks transport reactions and does not denote where reactions are occurring, i.e. within the cytosol, extracellular or other places. KEGG is also no longer freely downloadable. They still have a web page and REST API to access their data, but the user is not informed if they actively update this information or if it lags behind the paid subscribers version. The Model SEED database is another resource for genomes, metabolites, reactions and even full models(Overbeek, Disz, and Stevens, 2004). The Model SEED utilizes Rapid Annotation using Subsystems Technology (RAST) which performs the gene annotation and FBA modeling for you. RAST models can utilize genomes uploaded in FASTA or publicly available sequences in the Model SEED database. The big limitation to using Model SEED is that this data does not appear to be actively curated at this time and thus the possibility of inaccurate or incomplete data exists. When using RAST with Model SEED, another limitation is that during the gap filling step it adds a large number of low-confidence reactions in order to complete pathways. Finally, two manually created databases were created and curated by Dr. Niti Vanee and published by Dr. Bernhard Palsson (Vanee, 2013; Shlomi et al., 2008). These databases were built to address the missing transport reactions, lack of detail about reaction locations, and otherwise update and curate the missing pieces of data for KEGG and SEED.

1.2 Constraint-based Modeling

Constraint-based modeling is an approach that has been evolving since the 1980s (Fell and Small, 1986; Majewski and Domach, 1990). Initially, the approach was first shown to be viable when experimentally obtained metabolic fluxes and growth rates were shown to be consistent with computationally derived fluxes calculated from cellular objective functions (Savinell and B. O. Palsson, 1992; Schuster and Hilgetag, 1994). Then in the early 2000s when the ability to sequence whole genomes became more readily accomplished it became possible to link the genome directly to a constraint-based model. This link paved the way for using these models to predict experimental outcomes. For example, gene knockouts and changes in cellular behavior. As biology entered into the age of ’omic data it became possible to incorporate experimentally obtained transcriptomic, exomic, proteomic, and even metabolomic data into these models to further the ability to analyze and experiment *in silico*.

In general, constraint-based modeling works under the law of conservation of mass and that biomass growth and energy use can be used to predict metabolic fluxes for an organism (Schilling, Letscher, and B. O. Palsson, 2000; Schuster and Hilgetag, 1994). This is accomplished by first curating all the metabolites and reactions determined, or predicted, to be present in an organism. In the case of genome-scale metabolic networks, this is done by creating a stoichiometric matrix. The stoichiometric matrix is a versatile and consistent format present in constraint-based models that indicate the number of molecules used and created in reaction. Here we focus on constraint-based modeling for genome-scale metabolic reconstructions, it has also been used for signaling, transcriptional regulation and macromolecule synthesis (Papin and Bernhard O. Palsson, 2004; Li et al., 2009).

Compared to other modeling methods constraint-based modeling, in general, allows greater influence of metabolic networks for an organism and in a more realistic fashion. More specifically, an organism *in vivo* is subjected to physical, environmental, and physiochemical inhibitors and thus doesn't have an unlimited growth potential. By having the ability to apply these constraints makes the *in silico* models more accurate and also expands the ability to perform *in silico* experiments. Utilizing constraint-based modeling we are able to better determine the cellular behaviors of an organism when subjected to different external or internal influences. The end result of this is a series of reaction pathways represented as a flow chart or map that represents what an organism uses to sustain life and these pathways can often vary based on the specific constraints applied to the model.

1.2.1 Flux Balance Analysis

Flux Balance Analysis (FBA) is one such mathematical approach to modeling and analyzing the networks that make up an organism and is particularly common in genome-scale metabolic network reconstructions (Schuster and Hilgetag, 1994; Varma and B O Palsson, 1994; Thiele et al., 2009). FBA is a specific application of linear programming (LP) used to calculate and optimize the flow of metabolites over time through the biochemical reactions present in an organism to determine the steady-state flux distribution that maximizes the biomass yield. Given the stoichiometric matrix (S) and fluxes (v), the steady-state is represented as $Sv = 0$ and defines a system of linear equations. Next, to solve these equations we define an objective function, like biomass, and to predict the maximum growth rate we use $Z = c^T v$, where c is a vector of zeros with a value of 1 only in the reaction of interest. When we're using the biomass reaction, c has a value of one so we can represent this as:

$$Z = v_{\text{biomass}}$$

with parameters:

$$Sv + b^{src} - b^{esc} = 0$$

$$L \leq v \leq U$$

$$L^{src} \leq b^{src} \leq U^{src}$$

$$L^{esc} \leq b^{esc} \leq U^{esc}$$

where L and U define the lower and upper bounds for each reaction, and b^{src}, b^{esc} are the escape and source reactions specifically(Brooks et al., 2012). Finally we calculate the flux values that maximize Z.

1.2.2 Mixed Integer Linear Programming (MILP)

Mixed Integer Linear Programming (MILP) is another modeling method. MILP and LP are both general optimization modeling frameworks and have many applications outside of metabolic reconstructions. In comparing MILP and LP, MILP is designed to better incorporate and optimize the use of experimentally obtained data into the model as it lets you add integer restrictions variable values(Bordbar et al., 2014). This step helps improve model quality by attempting to reduce false positive and false negative values from experimental data(Vanee, 2013). A false positive is when a metabolite is predicted to be present but the reaction/gene is associated with producing the metabolite is not actually shown to be present based on experimental evidence. Similarly, a false negative is when a gene or reaction is incorrectly omitted from a model but experimental evidence shows that the associated gene and gene product are in fact present. These false values are believed to be caused by post-transcriptional regulation or alternative flux distributions, which are likely from isozymes and alternative pathways. In MetModel and Model SEED, MILP is used in FBA-Gap and GapFill which are two different algorithms designed to identify and correct reactions missing from pathways. As mentioned previously MILP can be an effective way to incorporate proteomic and other data into the pathway reconstructions (Shlomi et al., 2008).

In the case of MILP we have a problem expressed as:

$$maximize \quad cx + dy$$

with parameters:

$$Ax + By \leq b$$
$$x \in \mathbb{R}^n_+$$
$$y \in \mathbb{Z}^p_+$$

where $cx + dy$ is the objective function, $Ax + By \leq b$ are constraints, x and y are vectors of the decision variables(Brooks, 2005). We can now determine solutions for our objective function if they exist. It is possible for no solution or multiple solutions to exist, and the solution that provides the best objective function value is called the optimal solution. When we model using this type of function we look for these optimal solutions if they exist.

1.3 *In Silico* Bacteria Research

Modeling unicellular organisms *in silico* provides a number of benefits. It allows us to work and analyze with extremophiles and pathogens without expensive equipment or health hazards. It allows us to make predictions about the outcomes for in vivo or in vitro experiments before having to incur the temporal and fiscal costs associated with performing one (Langowski and Long, 2002). All of this together allows us to push research of treating and preventing diseases further by focusing and developing our understanding of virulence, pathogenesis and identifying new drug targets(Shlomi et al., 2008)(Nurputra et al., 2012). In industry different strains or even customized genomes can be tested using these methods and we can select a particular genome or strain of bacteria that provides optimal amounts of a given metabolite which can be collected for purposes like biofuels (Nogales, Gudmundsson, and Thiele, 2012).

1.3.1 Genome Annotation using ASGARD

Understanding the genes, their products and the metabolic reactions of *G. vaginalis* is crucial for researching the virulence, transmission, and therapeutics. We used the genomes of *G. vaginalis* strains obtained from NCBI and other sources, then use the Automated System for Gene Annotation and Metabolic Pathway Reconstruction Using General Sequence Databases (ASGARD) to determine open reading frames and annotate the genome(Alves and Buck, 2007).

ASGARD can take assembled sequences in FASTA file format and perform gene annotation and predicted metabolic pathways. The data provided by ASGARD can be regarded as a draft model, and this creates the first step to a high-quality metabolic model of our organism.

ASGARD creates these models by first determining the open reading frames within a genome by comparing it to annotated genomes stored in databases like NCBI Nucleotide and KEGG. Once the genes and their functions are determined it places them within the appropriate pathways. From here the model can be regarded as a "rough draft" as ASGARD has made an educated guess about the pathways and enzymes present based on translated nucleotide sequence homology only.

Using ASGARD thus allows one to take assembled genomic nucleotide sequences in FASTA file format and obtain gene annotation and predicted metabolic pathways. The data provided by ASGARD begins the search for an accurate metabolic model of our organism. We used both well-documented strains (i.e. strains that have already been annotated thoroughly) and novel strains. This "draft" model was then integrated with our MetModel where a series of scripts were used to integrate gene expression data, metabolic data and our other information to increase the accuracy and precision of the draft model.

1.3.2 Metabolic Pathway Reconstruction Using MetModel

For our purposes, ASGARD is just the first step and the model will undergo further revisions as it goes through the MetModel pipeline. The MetModel pipeline will gap fill pathways then use FBA to derive the reactions rates for optimal growth. It can then be used to build KGML maps of the reaction pathways and if available increase the accuracy and confidence we have in the metabolic reconstruction model by incorporating experimental data. mRNA expression data can be obtained from NCBI

Gene Expression Omnibus (GEO) and incorporated during this process in order to provide experimental data to support the analysis and solutions obtained from the FBA. Finally the model pathways were viewed and reviewed manually using KGML-ED(Klukas and Schreiber, 2007).

MetModel is Python library which can be used in a pipeline to gap-fill reaction pathways determined by ASGARD to then apply a constraint-based modeling approach. This modeling approach considers all of the potential biochemical reactions and then applies constraints in the same way that an organisms environment, physiochemical, regulatory and evolutionary sources would constrain its growth potential. Thus MetModel allows us to incorporate metabolic data with gene/reaction network, thermodynamics, gene regulation and other information. Using MetModel allows us to consider the states that an organism can and cannot achieve which gives us a broader view as to the factors that are involved in determining an organism's survivability, growth potential and even its ability to produce metabolites under various conditions and with greater accuracy than other modeling tools(Roberts et al., 2009).

In order for MetModel to perform these tasks it first converts the biochemical reaction network reconstruction into a mathematical form. To do this we went through three steps, the first is the analysis of the reactions within the network. They usually fall into three main categories like metabolic, regulatory, and signaling. Next, the data derived from this analysis is used to form the stoichiometric matrix. This stoichiometric matrix is the mathematical representation or map where the chemical constraints are applied to the model. Now that we had our mathematical representation of the organism's pathways, flux balance analysis (FBA) can be performed to generate a solution or solutions(Orth, Thiele, and B. Ø. Palsson, 2010). FBA calculates the flow of metabolites through the network, and this makes it possible to predict the production rates of metabolites, the growth rate of an organism, and analyze specific pathways and even predict experimental outcomes (Lee et al., 2005). Put another way these *in silico* models allow predictions of phenotypes given a set of genes and reactions. For example, we can perform an *in silico* knockout model or we can try to optimize gene products which is particularly useful for nitrogen-fixing bacteria used in biofuel production (Nogales, Gudmundsson, and Thiele, 2012). In the case of pathogens like *G. vaginalis* we can use MetModel to test out new drug designs, or better understand how it might infect and gain a foothold among the normal vaginal bacterial community.

Using this modeling approach considers all of the potential biochemical reactions and then applies constraints in the same way that an organism's environment, physiochemical, regulatory and evolutionary sources would constrain its growth potential. Thus MetModel allows us to incorporate metabolic data with gene/reaction network, thermodynamics, gene regulation and other constraints. This approach to modeling allows us to consider the states that an organism can and cannot achieve which gives us a broader view as to the factors that are involved in determining an organism's survivability, growth potential and even its ability to produce metabolites under various conditions.

1.4 Gardnerella

Gardnerella is a genus of bacteria for which *G. vaginalis* is presently the only known species. *G. vaginalis* is a clinically significant bacterium that can disrupt the normal vaginal flora and cause bacterial vaginosis

(BV). BV is a major medical problem, causing discomfort to millions of women every year and has been shown to cause complications for many pregnant women resulting in preterm labor and birth which may result in death or long-term health problems for the baby. Many patients with BV are asymptomatic but occasionally have yellow or gray discharge, irritation, or a foul odor. Diagnosing BV can be difficult especially if the patient is asymptomatic. Figure 1.1 shows a diagrammatic depiction of how these bacterial cells are identified once stained. *G. vaginalis* is not considered to be the single microbe inducing BV but rather a signal that the normal vaginal tract flora has been disrupted, thus paving the way for other anaerobes to work synergistically to reduce the protective, hydrogen peroxide producing *Lactobacillus* species that suppress the harmful bacteria from proliferating. Further, the *G. vaginalis* cells are so small they do not reliably show up as gram-positive and thus can be difficult to detect. Presently, the main treatment for patients with BV caused by *G. vaginalis* are antibiotics such as clindamycin or metronidazole.

Gardnerella vaginalis is a gram-variable anaerobic coccobacilli. It is a facultative anaerobe and can metabolize glucose under both aerobic and anaerobic conditions, and has a complex metabolism (Patterson et al., 2010). It is the sole member of the *Gardnerella* genus and is a small ($1.0\mu m$), non-motile and nonspore-forming bacterium. The *G. vaginalis* genome is a circular DNA and is without plasmids. Within *G. vaginalis* there are genetic variants that include both virulent and avirulent strains. It is considered to be a key component in the initiation and progression of BV (Schwebke, Muzny, and Josey, 2014). Models of the pathogenesis of BV suggest the virulent stains of *G. vaginalis* are usually transmitted through sexual intercourse and its virulence factors allow it to adhere

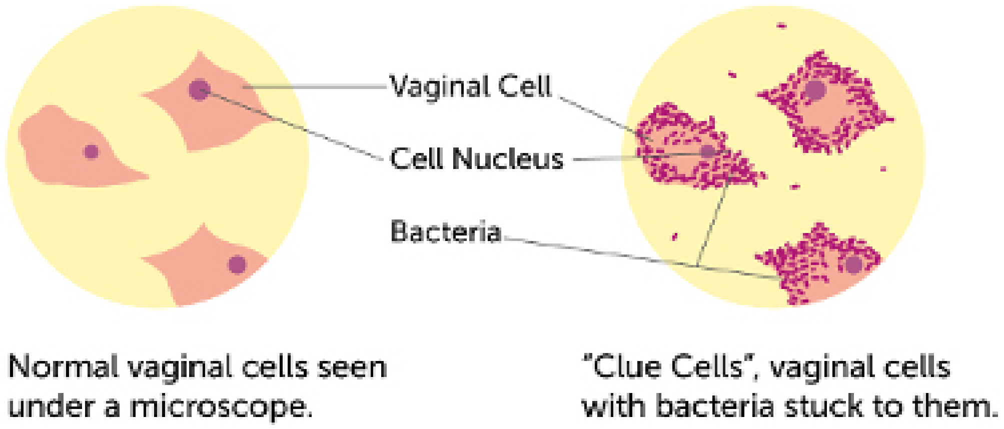
Vaginal Cell
Cell Nucleus
Bacteria
Normal vaginal cells seen under a microscope.
"Clue Cells", vaginal cells with bacteria stuck to them.

FIGURE 1.1: Artist's rendering of how a clinician is able to use microscopy to identify *G. vaginalis* cells that have infected vaginal tissue. Image reprinted with permission from: (*Bacterial Vaginosis | Center for Young Women's Health* 2016)

to vaginal epithelial tissue. Once attached to epithelial cells it creates a biofilm where a community of normally dormant vaginal anaerobes flourish. *Gardnerella vaginalis* also exhibits cytotoxic activities (Patterson et al., 2010). Once established this biofilm community then aggressively competes with microorganisms of the typical vaginal flora. For example, the predominant *Lactobacillus* populations that help regulate a healthy pH, and creates conditions for an overgrowth of *G. vaginalis* and its associated pathogenic anaerobes. This microfloral replacement results in the clinical symptoms associated with BV (odor, discomfort, itch etc.). Studying the pathogens' genes gives biochemical and metabolic information to interpret its cooperative and competitive interactions with its human host and co-occurring species, suggesting how it overgrows and out-competes the established healthy microflora. Understanding the etiology of the disease will hopefully give insights into the best methods to prevent and control it.

There are over 1,365 genes in the reference genome of *Gardnerella vaginalis ATCC 14019*. Not all of the *G. vaginalis* strains identified are virulent and more research is needed in order to understand the virulence factors. It does appear that *G. vaginalis* forms symbiotic relationships with other vaginal anaerobes that are normally dormant, and these relationships contribute to its success, resulting in symptoms and progression of BV(Gardner, 1983; Schwebke, Muzny, and Josey, 2014).

Chapter 2

Curating a Database for Metabolic Reconstructions

2.1 Reference Databases

Upon starting this project MetModel already used a comma separated file (CSV) that contained reactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG), the SEED database and contributions from Dr. Niti Vanee and published by Dr. Bernhard Palsson (Ogata et al., 1999; Overbeek, Disz, and Stevens, 2004; Vanee, 2009; Shlomi et al., 2008). However, despite these reactions all being in a single file, there was no way to relate the reactions to each other. It was apparent that there were duplicate reactions that were represented in different formats and that the overall process of looking up reactions could be improved by creating a standardized format for the reactions.

TABLE 2.1: An example of the data stored in new Compound Reference SQL Table.

KEGG ID	SEED ID	CHEBI ID	VANEE	PALSSON	Name
C00001	cpd00001	15377	H_2O	H_2O	Water
C00002	cpd00002	15422	ATP	ATP	ATP
C00003	cpd00003	13389	NAD+	NAD+	NAD
C00004	cpd00004	16908	NADH	NADH	NADH

2.1.1 Reference Database Collection and Clean Up

To collect and clean up the information housed in the Kyoto Encyclopedia of Genes and Genomes (KEGG), Chemical Entities of Biological Interest (ChEBI), and the SEED database, when SQL or CSV files were available they were downloaded, but often information needed to be scraped from these online sources(Ogata et al., 1999; Degtyarenko et al., 2008; Overbeek, Disz, and Stevens, 2004). Web scraping was performed using Scrapy (*Scrapy | A Fast and Powerful Scraping and Web Crawling Framework* 2016). Scrapy is a web scraping toolkit written in Python. Scrapy made it possible to download all of the information from these websites and simultaneously format it in a standardized way that we could then parse and load into a PostgreSQL database.

Loading all of this data into a SQL database made it possible to query this data simultaneously. Having all this data in a single place then allowed us to develop a Python pipeline to query each of these sources concurrently to return the identifiers for a given compound or reaction associated within each of these respective databases. This allowed for the creation of a standardized format and thus reduce duplicate information. For example, one of the biggest issues with the reactions is how the compounds are named. KEGG may refer to water as H_2O while SEED may actually refer to it as water. In another example, when water is donating a proton in a particular reaction some databases referred this as just H while in others H+,H_2O or even H_3O^+ even though the reaction was the same and clearly involved a single H (proton) being donated. With the methods described here we obtained the full set of metabolites and their associated information and we used pattern matching to automate the translation of these reactions into a standardized format using the KEGG identifiers (if available) for that given compound. The format took after the form of the KEGG identifiers like $C00001 + C00404 <=> C02174$ where C00001 represents H2O, C00404 represents polyphosphate and C02174 represents oligophosphate. If the compound was not found in the KEGG database but was present in others it was assigned a UNK000X identifier. Table 2.1 shows a sample of the results from the compound reference table.

Any reactions not automatically translated were flagged and reviewed manually. Once all of the compounds and reactions were in the same format we then quickly created a mapping of like equations. This mapping was stored in a PostgreSQL table so that we could quickly access relevant information in each database by retrieving its appropriate ID from the database. Table 2.2 shows a sample of the results from the reaction reference table.

Once completed, using Python and SQL statements, a quick and easy method to retrieve all of the relevant data and analysis resources from these pathway/genome databases was created. This rapid look-up helped us obtain and verify metabolic pathways and enzymes derived from experimental results published in the scientific literature. In particular, this is needed because unfortunately these databases are not always well maintained and information in any one source may be out of date or inaccurate.

Table 2.2: An example of the data stored in new Reaction Reference SQL Table.

KEGG ID	SEED ID	NITI	PALSSON	EC Number(s)
R01867	rxn09563	R_DHORD4	R_DHORD4	1.3.3.1
R04749	rxn03250	R_ECOAH2	R_ECOAH2	4.2.1.17|4.2.1.74
R00405	rxn00285	R_SUCOAS	R_SUCOAS	6.2.1.4|6.2.1.5
R03146	rxn10115	R_FDH3	R_FDH3	1.2.2.1

By using this method we kept up to date in order to best assign and verify the function for the majority of genes in selected genomes.

Chapter 3

From Sequence to Metabolic Reconstruction

ASGARD and MetModel are tools that perform genome annotation, and metabolic pathway reconstruction respectively. Both tools do their job well but it is not easy or intuitive to use them and especially not together. For starters, neither of these tools provide adequate documentation on how to use them or the specific file formats and data that they require to run. Next, they don't integrate well on their own so it was not possible to start from the raw nucleotide sequence data and build a metabolic model from there before our work. Now while the process is still not completely automated, that decision was actually intentional as it is helpful to be able to review the output from each step in the workflow in order to ensure nothing is missing and manually add or remove reactions or metabolites if need be. Figure 3-1 shows an overview of the pipeline and the steps that MetModel uses to perform FBA and generate the KGML reaction maps. Our specific contributions are highlighted in yellow, and the scoring function which was developed by Stephen Wunsch and then incorporated into the pipeline is highlighted in green(Wunsch, Stephen A., 2016).

3.1 ASGARD Parser

As mentioned before, ASGARD is a tool for determining open reading frames and then annotating genes, then uses this protein product information to determine the reactions present in the gene. It works by supplying a FASTA formatted file, and the output is a BED file which contains the enzyme commission numbers (EC) that were predicted to be present in the organism. This is a big step toward building a complete pathway model, but there were still missing steps before a model could be generated. First, a python script was developed and used to parse the EC numbers from the output, and searched within the new reference database for the associated reactions, pathways, and if possible genes/gene products associated with them. This information was then used to build the list of reactions needed to run MetModel.

3.2 MetModel Pipeline

Once the data from ASGARD was formatted properly it could now be used with the MetModel pipeline. In order to use MetModel originally you either had to create your own scripts and call the appropriate functions or for convenience four separate static scripts were written as example usages and each had to be tailored specifically to the model that was being run and each script performed a single step which needed to be executed independently and in order. As a part of this work all of the function calls, data, and information that was contained in these four separate scripts were incorporated into one Python script which created an user friendly, reusable software tool. The new script uses command line arguments to run different procedures and can be run dynamically without having to alter the code of the scripts themselves. Doing this created a semi-automated process in which a user can use to pause at each step if he or she wishes to manually intervene or review the files produced before completing the entirety of the MetModel FBA process.

This tool we created allows the user to also select which steps in the pathway to perform, the default being all four. Step 1 the MetModel pipeline adds transport reactions and if desired you can even attempt to build a model from this information. In step 2 we perform gap filling to complete the pathways and use FBA to determine the reaction rates and fluxes. In step 3, if experimental data is available it can be incorporated. In Step 4, the KGML maps are rendered. These models were then scored and validated using the scoring function implemented by Stephen Wunsch(Wunsch, Stephen A., 2016).

The scoring method is designed to give a relative confidence score in the pathways that were included in the final model. It is the result of a collaborative effor between Dr. Stephen Fong, Stephen Wunsch and myself and was ultimately implemented by Stephen Wunsch, a PSM Bioinformatics graduate student(Wunsch, Stephen A., 2016) The method he developed works by using BeautifulSoup, a Python library and framework for webscraping, similar to Scrapy. It takes an individual reaction within a network and uses BeautifulSoup to seek out publications that provide experimental evidence of that reaction within the pathway and organism. The more unique data it can discover the higher the score. It then outputs these scores on a scale of 1-10, 1 being the lowest and 10 being the highest. A score of 10 indicates that the gene, protein, and reaction all have at least one primary journal article supporting them that contains experimental evidence that explicitly shows that the gene products and reactions are present in the organism. The score decreases from there when evidence cannot be found for example, a score of 5 indicates that the gene-protein reaction (GPR) have been associated with multiple EC numbers but are without publications that provide direct experimental evidence to support them.

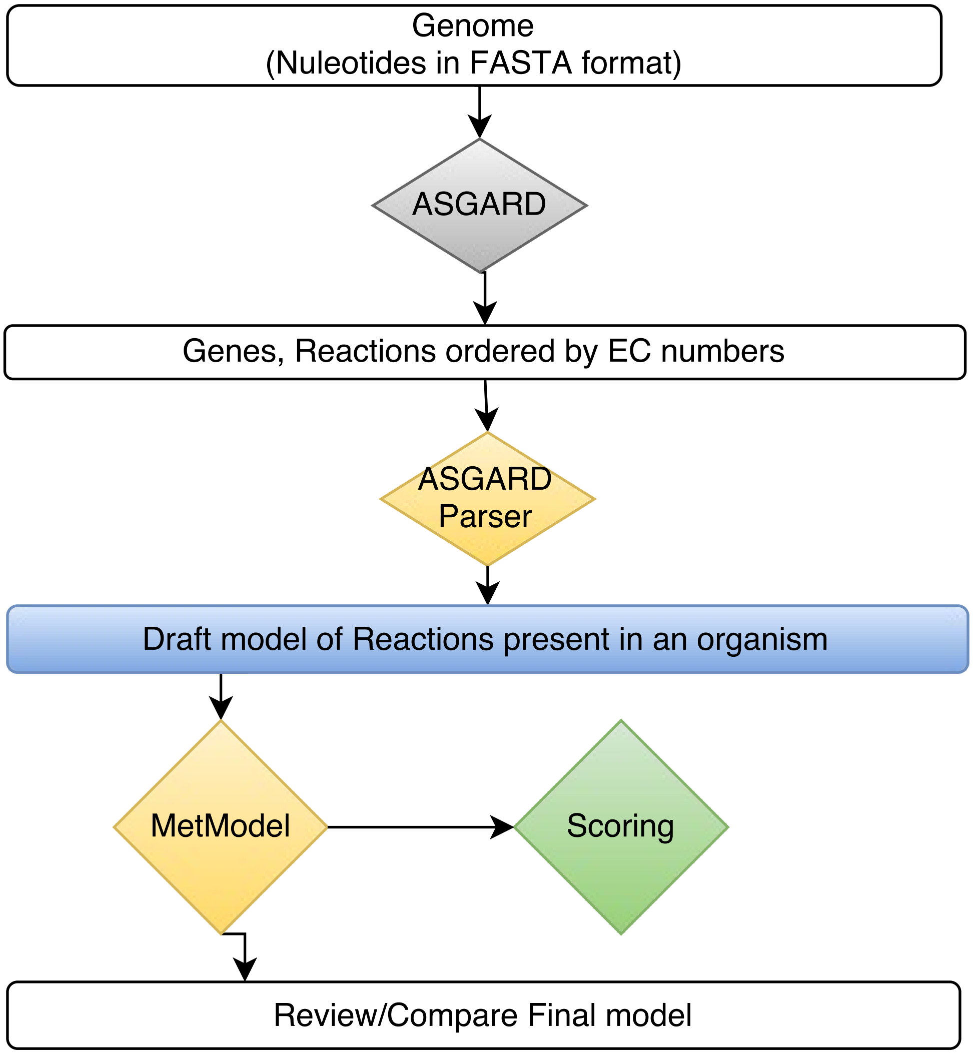
Genome
(Nuleotides in FASTA format)
ASGARD
Genes, Reactions ordered by EC numbers
ASGARD Parser
Draft model of Reactions present in an organism
MetModel
Scoring
Review/Compare Final model

Figure 3.1: This shows the workflow used to reconstruct a metabolic network starting from just the nucleotide sequence of an organism's genome. My specific contributions are highlighted in yellow, and the collaborative effort on the scoring mechanism is highlighted in Green.

Chapter 4

Comparison of Metabolism using Five *G. Vaginalis* Strains

4.1 *Gardnerella vaginalis* Nucleotide Sequences

The nucleotide sequences of the genomes from five different *Gardnerella vaginalis* strains: 5-1, 41V, 101, AMD, and ATCC 14019, were obtained from NCBI Nucleotide database. These nucleotide sequences in FASTA format were then uploaded to our computing cluster where they could be annotated and then reconstructed into models. Initial ASGARD annotations were also provided in Excel format containing ASGARD output from these five strains and others. This data was used for comparison against our ASGARD runs. We also downloaded the metabolic models generated using Model SEED for these same strains so that our results could be compared.

TABLE 4.1: Comparison of Gene Annotations from ASGARD, NCBI and Model SEED.

***G. vaginalis* Strain**	**ASGARD**	**NCBI**	**Model SEED**
5-1	2140	1271	345
AMD	2417	1190	339
101	1833	1150	329
ATCC 14019	1485	1366	380
41V	1210	1230	371

4.2 Genome Annotation using ASGARD

Understanding the genes, their products and the metabolic reactions of *G. vaginalis* is crucial for researching the virulence, transmission, and therapeutics. We used genomes of *G. vaginalis* strains obtained from NCBI and other sources, then used the Automated System for Gene Annotation and Metabolic Pathway Reconstruction Using General Sequence Databases (ASGARD) to determine open reading frames and annotate the genome(Alves and Buck, 2007).

By using ASGARD we were able to take the assembled nucleotide sequences, obtained from NCBI's nucleotide database, in FASTA file format and obtain gene annotation and predicted metabolic pathways. The data provided by ASGARD is a set of reactions for a given pathway, determined by the genes which were present during the annotation. We regarded this data as a draft model for the reactions present in each of the *Gardnerella* strains. Each of these draft models was then run through our ASGARD parser script, which extracted the EC numbers in each of the pathways that were determined by ASGARD to be present. The EC numbers were then used to obtained the specific reactions and when possible associated genes, by querying our reference database. This information was then put into the appropriate text format so that it could move on to the next step of being put through the MetModel pipeline where we would integrate gene expression data, metabolic data and our other information to increase the accuracy of the model.

To validate the gene annotation performed by ASGARD we compared the Gene Feature Format (GFF) file created by ASGARD and the GFF files obtained from NCBI and the Model SEED table containing genes and their reactions. The GFF files contain genes and their coordinates and we compared them both by looking at the number of genes and the name of the gene. We removed redundant genes within the ASGARD GFF and NCBI GFF files before performing this comparison. Table 4.1 shows the results of this comparison. Overall, ASGARD showed an average of 72% similarity in the genes determined to be present between the strains when compared to the genes present in the NCBI annotations for each strain. Although, there were was a strong deviation for the AMD strain which was only 49% similar and ASGARD determined a significantly higher amount of genes found compared to the number present in the reference strain. On the contrary, the 41V strain was 98% similar. In all cases, ASGARD determined a larger number of genes when compared to the number of genes predicted by Model SEED. This difference appeared to be due to Model SEED only regarding genes in the PATRIC database that have EC numbers attached to them (Devoid et al., 2013).

It is clear that the ASGARD algorithm is also more greedy than the Model SEED algorithm when it comes to gene and reaction pathway determination. However, since these strains lacked experimental evidence for transcriptomic or proteomic data it is unknown if the accuracy of ASGARD to determine open reading frames and predict the genes present in an organism is better or worse than Model SEED's annotation process. This increased amount of genes, and therefore pathways, as predicted initially by ASGARD and used as a starting point my MetModel, did eliminate the need for gapfilling during pathway reconstruction, which is a positive outcome and could indicate the ASGARD is more thorough and accurate when annotating a genome.

4.3 Metabolic Pathway Reconstruction Using MetModel

Each of the genes and reaction sets for all of the strains of *Gardnerella* obtained from ASGARD, and parsed out into the appropriate format expected by MetModel were then run through the MetModel pipeline using the new MetModel tool. Our MetModel tool allows a user to start from a file that contains gene-reaction products and run through four different steps to take a set of reactions and reconstruct the individual reaction pathways in order to model an organism. For all of the strains of *Gardnerella* we ran all through the steps of the MetModel script, excluding step 3 as no experimental proteomic data was available for any of the individual strains. The MetModel pipeline then allowed us to use the set of genes and reaction pathways determined by ASGARD to then apply the FBA constraint-based modeling approach.

The reconstructions created by MetModel were then compared to models available in the Model SEED database. We found that on average our reconstructions had 474 more reactions than the Model SEED reconstructions. Another major difference to note is that during the gap-filling step in the MetModel pipeline no reactions needed to be added in order to complete pathways. Of course as previously mentioned we did add transports and escapes during the first iteration in the MetModel pipeline. It appears that MetModel ended up with more reactions because the reaction data parsed ASGARD had a much higher number of genes and reactions. While some of these reactions were removed by MetModel a significant amount stayed and thus increased the number of reactions compared to

Model SEED. Comparing MetModel reconstructions to the Model SEED reconstructions it at first seemed odd there was a difference of over 400 reactions, but when looking at other models, for example, *Escherichia coli K-12 MG1655* it contains 1366 genes and 2251 reactions in MetModel reconstruction and in the Orth et al. 2011 published model while in Model SEED it contains only 1132 genes and 1632 reactions. Further, the *E. coli* only had transports and escapes added prior to the gap-filling step (GapFill) in MetModel while in Model SEED 38 reactions were added(Brooks et al., 2012).

One of the principle reasons for the differences in the number of genes present from each of the annotation sources is the algorithms used to determine the genes present. ASGARD uses a greedy algorithm and it appears it could be overestimating the number of genes present. NCBI, on the other hand, uses information uploaded by its users so depending on the methods used to annotate the organism's genome the accuracy can vary. Further, the NCBI data is not always well curated so it is also possible that some older methods and technologies were used to sequence and annotate these organisms which could again affect the accuracy of the sequences and accuracy of the gene annotations. Finally, Model Seed appears to be only including genes that are present in the PATRIC database and have known enzymes catalog identifiers attached to them. While this approach does ensure the genes predicted to be present have a high degree of experimental and literature support it is likely missing out on a lot of genes whose functions have limited evidence available but are, in fact, present in the organism.

Table 4.2: Comparing the number and types of Reactions in MetModel vs Model SEED.

***G. vaginalis* Strain**	**MetModel total**	**Model SEED total**
5-1	1217	761
AMD	1248	744
101	1220	760
ATCC 14019	1232	769
41V	1235	745

Table 4.3: The Average Confidence scores of the five strains.

***G. vaginalis* Strain**	**Score**
5-1	1.92
AMD	1.84
101	1.97
ATCC 14019	6.89
41V	1.94

4.4 MetModel Validation and Scoring

First, we compared the draft models from ASGARD to each other. We found that ASGARD determined 153 pathways in each of the strains, consisting of an average of 1802 reactions in total. We also compared these individual models against a previous ASGARD run after as we were utilizing updated reference databases from UniProt. This comparison revealed that there was no difference between our ASGARD pathway data and the previous version. This data was then formatted into the appropriate format and the MetModel pipeline was used without data integration. The MetModel pipeline added an average of 22 sources, 2 escapes and during the FBA-GAP no reactions were added. Overall the models had an average of 1230 reactions, and the reaction sets present in each given pathway were highly similar >85%. This high degree of similarity supports the results from the gene annotations from ASGARD where both the number of genes, types of genes and the initial pathway predictions were very similar as well.

Once we completed the models for each of these strains we then used the scoring function to determine relative confidence scores for each of the reactions and averaged them to produce an overall score for each individual strain model. The results shown in Table 4.3 demonstrate that with the exception of ATCC 14019 there was very minimal experimental data about reactions, pathways and gene products available for these *Gardnerella* strains. These scores of 2 or less indicate that there is no evidence in PubMed that supports the presence of the gene to protein to reaction association (GPR), with the exception of strain ATCC 14019 (with a score of 6.89) which has published evidence of the GPRs associated with its model. First, these results indicate that there is a clear lack of evidence supporting the reaction pathways determined by the pipeline to be present in the model. Thus it makes it difficult to say confidently that for these given strains of *G. vaginalis* have a high degree of accuracy as it is unknown if these GPRs are truly present in these organisms. Further, these results also indicated a problem with the current automated scoring system. The automated scoring system is designed to look for GPRs that have KEGG IDs for genes. In all the strains except ATCC 14019 no KEGG ID was given for the GPR as these organisms do not exist in KEGG, and thus the results are likely skewed.

Chapter 5

Future Directions

Here we have developed a semi-automated process for taking the nucleotide sequence of an organism's genome to reconstruction it's metabolic reaction networks. From there using our in-house developed scoring function we were able to assign a confidence score to help determine the quality of the reactions present in the model. While our results validate this process there are a few things that need more research and development.

Since there was no expression data available for these *G. vaginalis* strains if expression level data becomes publicly available it would be constructive to rebuild these models and incorporate that data. By incorporating experimental data the MetModel results will more accurately represent the organisms pathways. It would also be useful to determine the KEGG ID GPRs for these organisms, even if experimental data is unavailable at least by similarity, it could be possible to lend confidence to the models by relating known genes within the reference strains of *Gardnerella.*

Next, it would be useful to further improve the scoring process to return more information about the publications found. For example, the scoring function only does a single search for the GPR based on the information about the GPR in KEGG. It would be beneficial if it could also return data based on EC or even gene functional type in the event the organism is not a direct match. Further, the function presently does not return the dates or methods of the relevant publications and this could help improve confidence as more recent papers likely may have a greater degree of accuracy and precision.

Bibliography

Alves, João M. P. and Gregory A. Buck (2007). “Automated System for Gene Annotation and Metabolic Pathway Reconstruction Using General Sequence Databases”. en. In: *Chemistry & Biodiversity* 4.11, pp. 2593–2602. ISSN: 1612-1880. DOI: 10.1002/cbdv.200790212. URL: http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200790212/abstract (visited on 09/04/2015).

Bacterial Vaginosis / Center for Young Women’s Health (2016). URL: http://youngwomenshealth.org/2012/09/21/bacterial-vaginosis/ (visited on 09/01/2016).

Bordbar, Aarash et al. (2014). “Constraint-based models predict metabolic and associated cellular functions”. en. In: *Nature Reviews Genetics* 15.2, pp. 107–120. ISSN: 1471-0056. DOI: 10.1038/nrg3643. URL: http://www.nature.com.proxy.library.vcu.edu/nrg/journal/v15/n2/full/nrg3643.html (visited on 08/06/2016).

Brooks, J. Paul (2005). “Solving a mixed-integer programming formulation of a classification model with misclassification limits”. In: URL: https://smartech.gatech.edu/handle/1853/7473 (visited on 08/25/2016).

Brooks, J. Paul et al. (2012). “Gap Detection for Genome-Scale Constraint-Based Models”. In: *Advances in Bioinformatics* 2012. ISSN: 1687-8027.

DOI: 10.1155/2012/323472. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444828/ (visited on 08/25/2016).

Degtyarenko, Kirill et al. (2008). "ChEBI: a database and ontology for chemical entities of biological interest". en. In: *Nucleic Acids Research* 36.suppl 1, pp. D344–D350. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkm791. URL: http://nar.oxfordjournals.org/content/36/suppl_1/D344 (visited on 07/12/2016).

Devoid, Scott et al. (2013). "Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED". eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 985, pp. 17–45. ISSN: 1940-6029. DOI: 10.1007/978-1-62703-299-5_2.

Fell, D. A. and J. R. Small (1986). "Fat synthesis in adipose tissue. An examination of stoichiometric constraints". eng. In: *The Biochemical Journal* 238.3, pp. 781–786. ISSN: 0264-6021.

Gardner, H. L. (1983). "Pathogenicity of Gardnerella vaginalis (Haemophilus vaginalis)". eng. In: *Scandinavian Journal of Infectious Diseases. Supplementum* 40, pp. 37–40. ISSN: 0300-8878.

Keseler, Ingrid M. et al. (2013). "EcoCyc: fusing model organism databases with systems biology". en. In: *Nucleic Acids Research* 41.D1, pp. D605–D612. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gks1027. URL: http://nar.oxfordjournals.org/content/41/D1/D605 (visited on 08/28/2015).

Klukas, Christian and Falk Schreiber (2007). "Dynamic exploration and editing of KEGG pathway diagrams". en. In: *Bioinformatics* 23.3, pp. 344–350. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btl611. URL: http://bioinformatics.oxfordjournals.org/content/23/3/344 (visited on 07/12/2016).

Langowski, Jan and Anthony Long (2002). “Computer systems for the prediction of xenobiotic metabolism”. In: *Advanced Drug Delivery Reviews* 54.3, pp. 407–415. ISSN: 0169-409X. DOI: 10.1016/S0169-409X(02)00011-X. URL: http://www.sciencedirect.com/science/article/pii/S0169409X0200011X (visited on 03/02/2014).

Lee, Dong-Yup et al. (2005). “Complementary identification of multiple flux distributions and multiple metabolic pathways”. In: *Metabolic Engineering* 7.3, pp. 182–200. ISSN: 1096-7176. DOI: 10.1016/j.ymben.2005.02.002. URL: http://www.sciencedirect.com/science/article/pii/S1096717605000194 (visited on 03/02/2014).

Li, Fan et al. (2009). “Identification of Potential Pathway Mediation Targets in Toll-like Receptor Signaling”. In: *PLOS Computational Biology* 5.2, e1000292. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000292. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000292 (visited on 12/13/2016).

Majewski, R. A. and M. M. Domach (1990). “Simple constrained-optimization view of acetate overflow in E. coli”. eng. In: *Biotechnology and Bioengineering* 35.7, pp. 732–738. ISSN: 0006-3592. DOI: 10.1002/bit.260350711.

Nogales, Juan, Steinn Gudmundsson, and Ines Thiele (2012). “An in silico re-design of the metabolism in Thermotoga maritima for increased biohydrogen production”. In: *International Journal of Hydrogen Energy* 37.17, pp. 12205–12218. ISSN: 0360-3199. DOI: 10.1016/j.ijhydene.2012.06.032. URL: http://www.sciencedirect.com/science/article/pii/S0360319912013791 (visited on 03/02/2014).

Nurputra, Dian K. et al. (2012). "Paramyotonia congenita: From clinical diagnosis to in silico protein modeling analysis". en. In: *Pediatrics International* 54.5, pp. 602–612. ISSN: 1442-200X. DOI: 10.1111/j.1442-200X.2012.03646.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1442-200X.2012.03646.x/abstract (visited on 07/12/2016).

Ogata, H et al. (1999). "KEGG: Kyoto Encyclopedia of Genes and Genomes." In: *Nucleic Acids Research* 27.1, pp. 29–34. ISSN: 0305-1048. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC148090/ (visited on 07/12/2016).

Orth, Jeffrey D., Ines Thiele, and Bernhard Ø Palsson (2010). "What is flux balance analysis?" en. In: *Nature Biotechnology* 28.3, pp. 245–248. ISSN: 1087-0156. DOI: 10.1038/nbt.1614. URL: http://www.nature.com/nbt/journal/v28/n3/full/nbt.1614.html (visited on 09/17/2015).

Overbeek, Ross, Terry Disz, and Rick Stevens (2004). "The SEED: A Peer-to-Peer Environment for Genome Annotation". In: *Communications of the ACM* 47.11, pp. 46–51. ISSN: 00010782. URL: http://proxy.library.vcu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,cookie,uid&db=iih&AN=14958068&site=ehost-live&scope=site (visited on 07/12/2016).

Papin, Jason A. and Bernhard O. Palsson (2004). "The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis". eng. In: *Biophysical Journal* 87.1, pp. 37–46. ISSN: 0006-3495. DOI: 10.1529/biophysj.103.029884.

Patterson, Jennifer L. et al. (2010). "Analysis of adherence, biofilm formation and cytotoxicity suggests a greater virulence potential of Gardnerella vaginalis relative to other bacterial-vaginosis-associated anaerobes". In: *Microbiology* 156.2, pp. 392–399. DOI: 10.1099/mic.0.034280-0. URL: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.034280-0 (visited on 08/13/2016).

Roberts, Seth B. et al. (2009). "Proteomic and network analysis characterize stage-specific metabolism in Trypanosoma cruzi". In: *BMC Systems Biology* 3, p. 52. ISSN: 1752-0509. DOI: 10.1186/1752-0509-3-52. URL: http://dx.doi.org/10.1186/1752-0509-3-52 (visited on 09/01/2016).

Savinell, J. M. and B. O. Palsson (1992). "Optimal selection of metabolic fluxes for in vivo measurement. II. Application to Escherichia coli and hybridoma cell metabolism". eng. In: *Journal of Theoretical Biology* 155.2, pp. 215–242. ISSN: 0022-5193.

Schilling, C. H., D. Letscher, and B. O. Palsson (2000). "Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective". eng. In: *Journal of Theoretical Biology* 203.3, pp. 229–248. ISSN: 0022-5193. DOI: 10.1006/jtbi.2000.1073.

Schuster, Stefan and Claus Hilgetag (1994). "On elementary flux modes in biochemical reaction systems at steady state". In: *Journal of Biological Systems* 02.02, pp. 165–182. ISSN: 0218-3390. DOI: 10.1142/S0218339094000131. URL: http://www.worldscientific.com/doi/abs/10.1142/S0218339094000131 (visited on 12/13/2016).

Schwebke, Jane R., Christina A. Muzny, and William E. Josey (2014). "Role of Gardnerella vaginalis in the Pathogenesis of Bacterial Vaginosis: A Conceptual Model". en. In: *Journal of Infectious Diseases* 210.3, pp. 338–343. ISSN: 0022-1899, 1537-6613. DOI: 10.1093/infdis/jiu089. URL: http://jid.oxfordjournals.org.proxy.library.vcu.edu/content/210/3/338 (visited on 08/13/2016).

Scrapy | A Fast and Powerful Scraping and Web Crawling Framework (2016). URL: http://scrapy.org/ (visited on 07/12/2016).

Shlomi, Tomer et al. (2008). "Network-based prediction of human tissue-specific metabolism". en. In: *Nature Biotechnology* 26.9, pp. 1003–1010. ISSN: 1087-0156. DOI: 10.1038/nbt.1487. URL: http://www.nature.com.proxy.library.vcu.edu/nbt/journal/v26/n9/full/nbt.1487.html (visited on 07/12/2016).

Thiele, Ines et al. (2009). "Genome-Scale Reconstruction of Escherichia coli's Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization". In: *PLOS Computational Biology* 5.3, e1000312. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000312. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000312 (visited on 12/13/2016).

Vanee, Niti (2009). "The Genome Scale Metabolic Model of Cryptosporidium hominis: iNV209". In: *Theses and Dissertations*. URL: http://scholarscompass.vcu.edu/etd/1909.

— (2013). "HIGH THROUGHPUT DATA FRAMEWORK BASED CHARACTERIZATION AND EVALUATIONS OF THERMOBIFIDA FUSCA FOR INDUSTRIAL APPLICATIONS". In: *Theses and Dissertations*. URL: http://scholarscompass.vcu.edu/etd/3274.

Varma, A and B O Palsson (1994). “Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.” In: *Applied and Environmental Microbiology* 60.10, pp. 3724–3731. ISSN: 0099-2240. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC201879/ (visited on 12/13/2016).

Wunsch, Stephen A. (2016). *Scoring Method for Genome-Scale Metabolic Network Reconstructions.*

Appendix A

Users Guide - Building Genome Scale Metabolic Reconstructions

This will serve as a guide on how to actually execute the steps required to go from a nucleotide FASTA file to the complete KGML pathway maps for a given organism. This is targeted at VCU faculty, students, and staff as it will refer to specific locations and servers housed in the CHPC.

A.1 Using ASGARD

Asgard is installed on the distributed computing cluster called Godel. It makes use of Grid Engine to distribute the various Blast processes and other jobs to different nodes. Since it is installed on Godel you first have to have an account there. If you do not already have access to Godel ask your advisor for how you can go about obtaining one. For the rest of the guide we will assume that you have access to Godel via SSH/SFTP (remember off campus will require VPN access as well) and proper permissions to access the ASGARD and BLAST executables.

Start by uploading the FASTA file you wish to run through ASGARD, then login to Godel and enter the directory where your sequence is stored.

I highly recommend you have this sequence file alone and in its own appropriately named directory as it will make your life easier later on. Once in the directory you can run the following command to queue the ASGARD jobs:

```
/usr/global/blp/bin/asgard -i YOURSEQUENCE.fasta
-p blastx -n 20 -d /gpfs_fs/data/refdb/asgardDB/UniRef100
-d /gpfs_fs/data/refdb/asgardDB/KEGG
-f /gpfs_fs/data/refdb/asgardDB/uniref100.fasta.gz
-f /gpfs_fs/data/refdb/asgardDB/genes.pep.gz
-l /gpfs_fs/data/refdb/asgardDB
```

Where **/usr/global/blp/bin/asgard** is the location of the ASGARD executable, **-i** is the flag for your FASTA file, **-p** is which blast program to use (generally blastx but consult the NCBI Blast documentation if you are unsure), **-n** is the number of nodes to use, **-d** specifies the locations of the protein databases, while **-f** specifies the FASTA files that correspond to the databases specified in the -d command, and finally **-l** is the location of the mapping files.

Once ASGARD completes successfully, usually within a few hours, you will find a number of new files present in the directory where you stored your sequence. I'm going to focus on the four that are of interest in relation to create metabolic reconstructions. These five files will be named YOURSEQUENCE.fasta but have the extensions: .gff, .path_rec, .paths, .paths.detail, again where YOURSEQUENCE.fasta is the name if your FASTA file given to ASGARD. There are usually two files with the extension GFF which are the General Feature Format (GFF) files that contain information about the open reading frames identified by ASGARD. Next, the path files contain summary or detailed information about the genes implicated in pathways, and the pathways that were matched based on those genes. It is worthwhile at this time to review and make sure you understand what the output of these files are before continuing, but once satisfied run:

```
python asgard_parse.py YOURSEQUENCE.fasta.paths
```

and it will generate a single file that contains the information required to run the MetModel pipeline.

A.2 Using MetModel

Since MetModel is a Python Library it may be difficult to setup, luckily it is already installed on Dr. Brooks's server as well as godel. There aren't a lot of prerequisites for MetModel but you will need install Gurobi if it is not installed and you will need a license for Gurobi (even if it is already installed). Once you have Gurobi installed you can clone git repository hosted on GitHub: Met-Modeling on GitHub. In any case, you just need to make sure that the install locations are provided to your PYTHONPATH environmental variable. Contact a system administrator if you are unsure how to do this yourself. From here we will assume you can `import metmodel` from within Python. The rest of this guide assumes you are working with ASGARD data and are already within the working directory where you have your asgard_parse output. Once you have cloned the GitHub or have access to the MetModel Python library in your Python path and have downloaded the met_model.py script, you can now run the four steps in our metabolic reconstruction pipeline by typing:

```
python met_model.py -i YOURSEQUENCE.txt -t TXT -x exchanges
-b biomass -ndi
```

If you have metabolomic or gene expression data available you can specify

-m or **-d** for each respectively and **omit the -ndi** flag. For more information you can also run `python met_model.py -h` for a list of all the available options.

Once met_model.py is invoked it will run a step and then pause, asking you to continue, while paused it is possible to modify and view files as your see fit. Then once complete you will be able to view the KGML pathway maps using a KGML viewer of your choice.

VITA

Shaun Norris was born on June 27, 1986, in Falls Church, Virginia and is an American citizen. Shaun received his high school diploma from Commonwealth Academy in Alexandria, Virginia. He completed his Bachelors of Science from VCU Life Sciences at Virginia Commonwealth University, Richmond, Virginia in 2016, majoring in Bioinformatics.