Pseudomonas & Burkholderia Genome Databases

Student's Name
Co-op Course Number

Brinkman Lab

Simon Fraser University

Department of Molecular Biology and Biochemistry

Letter of Transmittal

To whom it may concern:

This report was prepared for Dr. Fiona Brinkman, an associate professor in the Department of Molecular Biology and Biochemistry at Simon Fraser University. It covers the project work that I completed during a co-op work term under Dr. Brinkman's supervision.

This report is to be used as a reference for any future employees or students that may use or modify the work that I completed. The content of this report is not confidential.

The focus of this report is on two databases: *Pseudomonas* Genome Database, and *Burkholderia* Genome Database. The *Burkholderia* database was created as my project using the *Pseudomonas* database as a template. The report covers the details of server set up and implementation of new features that were included in order to improve the quality and usage of the databases. The work was completed over 8 months, from May, 2007, to December, 2007.

Recommendations for future work are also included at the end of this report for future students or employees that may work on this project.

Sincerely,

Student's Name

Science Co-op Program
University of British Columbia
Department of Computer Science

Pseudomonas & Burkholderia Genome Databases

Student's Name

Work Term Period

Brinkman Lab
Simon Fraser University
Department of Molecular Biology and Biochemistry

Summary

Cystic fibrosis is one of the most common fatal genetic diseases affecting young Canadians and Americans today. Many of these patients have compromised immune systems and have difficulty fighting off infections. In particular, these patients are very susceptible to *Pseudomonas aeruginosa* and *Burkholderia cenocepacia* infections that sometimes have fatal results. Much research is being done to study the genomic properties of these organisms to understand their mode of pathogenesis (the mechanism and all contributing factors and interrelated processes that cause disease). So by understanding these mechanisms, researchers can develop new therapeutic drugs to clear the bacteria from infected patients as well as new vaccines to prevent infections altogether.

In order to accomplish this goal, the *Pseudomonas* Genome Database was developed in 2001. Since its release, it has been used by researchers all over the world to analyze and study the genomic properties of *Pseudomonas* species. Using this database as a template, the *Burkholderia* Genome Database was released in October, 2007. It provides similar tools for researchers studying the genomes of *Burkholderia* species.

Furthermore, additional tools and features were implemented in both databases to provide researchers with access to more comparative analysis capabilities. This allows researchers to not only access the data that they want, but to use a variety of methods to analyze it and determine the underlying processes that are leading to bacterial pathogenesis. Since the code for this project is freely available to the public, these databases provide a model for other scientists to follow in studying the genomic properties of other organisms.

Table of Contents

List of Figures

List of Tables

1 Introduction

"Cystic fibrosis is the most common, fatal genetic disease affecting young Canadians" (Canadian Cystic Fibrosis Foundation, 2007). It predominately affects the lungs and digestive system, causing difficulties in breathing and intake of nutrients. It also affects the individual's ability to fight disease. Their immune systems become very weak and they are increasingly susceptible to infections that may be difficult or even impossible to treat. Currently, there is no cure for this genetic disorder, but there is much research going on to find a cure.

One of the most common pathogens that infects cystic fibrosis patients is *Pseudomonas aeruginosa*, an environmentally versatile Gram-negative bacterium with a fairly large genome (6.3 Mega-base pairs). This bacterium causes disease, or sometimes even death, in immune-compromised individuals by producing toxins that may cause tissue damage or by interfering with the proper function of the immune system. This species is noted for its resistance to antibiotics, so treatment to infections can be quite difficult.

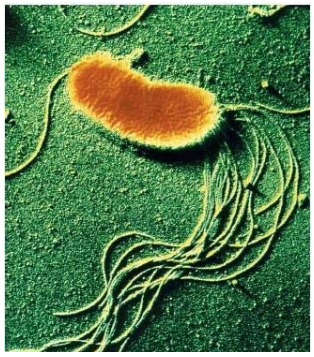
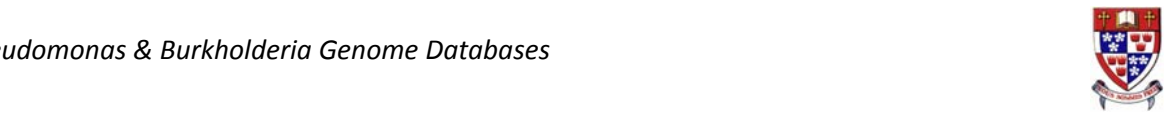


Figure 1. *Pseudomonas fluorescens*.

In order to facilitate research on the *Pseudomonas* genome to find a cure for the devastating affect it may have on these persons, the *Pseudomonas* Genome Database was released in 2001 (Winsor GL, 2005). With this resource, researchers across the globe have been able to access information about this bacterium and continually update the genome annotation based on their own specific projects. This community approach to updating allows researchers to access the most recent annotations of the *P. aeruginosa* genome. Some other features of this database include sequence searches, sequence alignments, GBrowse views of genes/proteins, ortholog pairs and



homology with humans. Overall, it provides a myriad of resources to the fingertips of those searching for clues into developing preventative techniques for these infections.

Figure 2. *Burkholderia cenocepacia* HI2424.

Another fairly infectious group of bacteria that colonizes cystic fibrosis patients are the species of the *Burkholderia cepacia* complex, most commonly *Burkholderia cenocepacia*. These species are also very environmentally versatile and have fairly large genomes. Though they are not as resistant to antibiotics as *Pseudomonas aeruginosa*, *Burkholderia cenocepacia* infections are also very difficult to treat with antibiotics. The *Burkholderia* Genome Database was designed using the *Pseudomonas* Genome Database as a template, in order to provide researchers with the same tools to study a different set of bacterial species.

Both the *Pseudomonas* Genome Database and the newly released *Burkholderia* Genome Database have been developed in order to facilitate research so that researchers can understand the basic biology of these organisms and their modes of infection. Overall, it is very important to study the genomes of these pathogens in order to find possible virulence factors, new drug candidates or even vaccine targets so that in the future, infections can be treated more easily or prevented altogether.

This report will cover aspects of the set up the *Burkholderia* Genome Database and all new tools and features that have been implemented in both databases. It will also describe how some of these methods can be applied in understanding more about these particular bacterial species.

2 Methods and Materials

Figure 3. Apache Tomcat

The web servers used for this project are running Apache Web Server 2.0 and Apache Tomcat 5.5.4, with SuSe Linux 10.2 operating system. Both websites are able to run simultaneously on a single server using virtual hosting technology. In total there are 2 servers, one used to host the live websites, and the other used as a developmental machine.

MySQL databases were used to store all of the data for the species available on these websites. All data for the *Pseudomonas* species was stored in a single database, while data for the *Burkholderia* species was stored in a separate database. Additional databases were created as required by other parts of the website, which will be covered in later sections of this report.

All pages of the website were written in Java Server Pages (JSP), which uses Java classes in the background to generate dynamic content by querying the databases. Some additional Perl scripts were also used to parse data files. There was no integrated development environment used, so all edits and changes to web pages were made through Unix text editors such as Emacs and vi.

Perl scripts were used throughout the term to perform analyses on the genomes and to load data into the databases. Considering that many of the analyses performed would take a very long time to run on a single computer, 'Buster the cluster' (a cluster of 60 CPU nodes), was used to parallelize jobs to decrease the time taken for each job to run. This dramatically cut down the time spent on running analyses from weeks to a matter of days.

3 Database Setup

This first part of my work term was spent on setting up the *Burkholderia* Genome Database (www.burkholderia.com) using the *Pseudomonas* Genome Database (www.pseudomonas.com) as a template. First of all, I will discuss some of the main features and tools provided through this website for researchers to study the genomes of *Pseudomonas* species.

3.1 Pseudomonas Genome Database Features

First of all, the *Pseudomonas* Genome Database provides robust, continually updated information about *Pseudomonas* genomes. All genomes (see

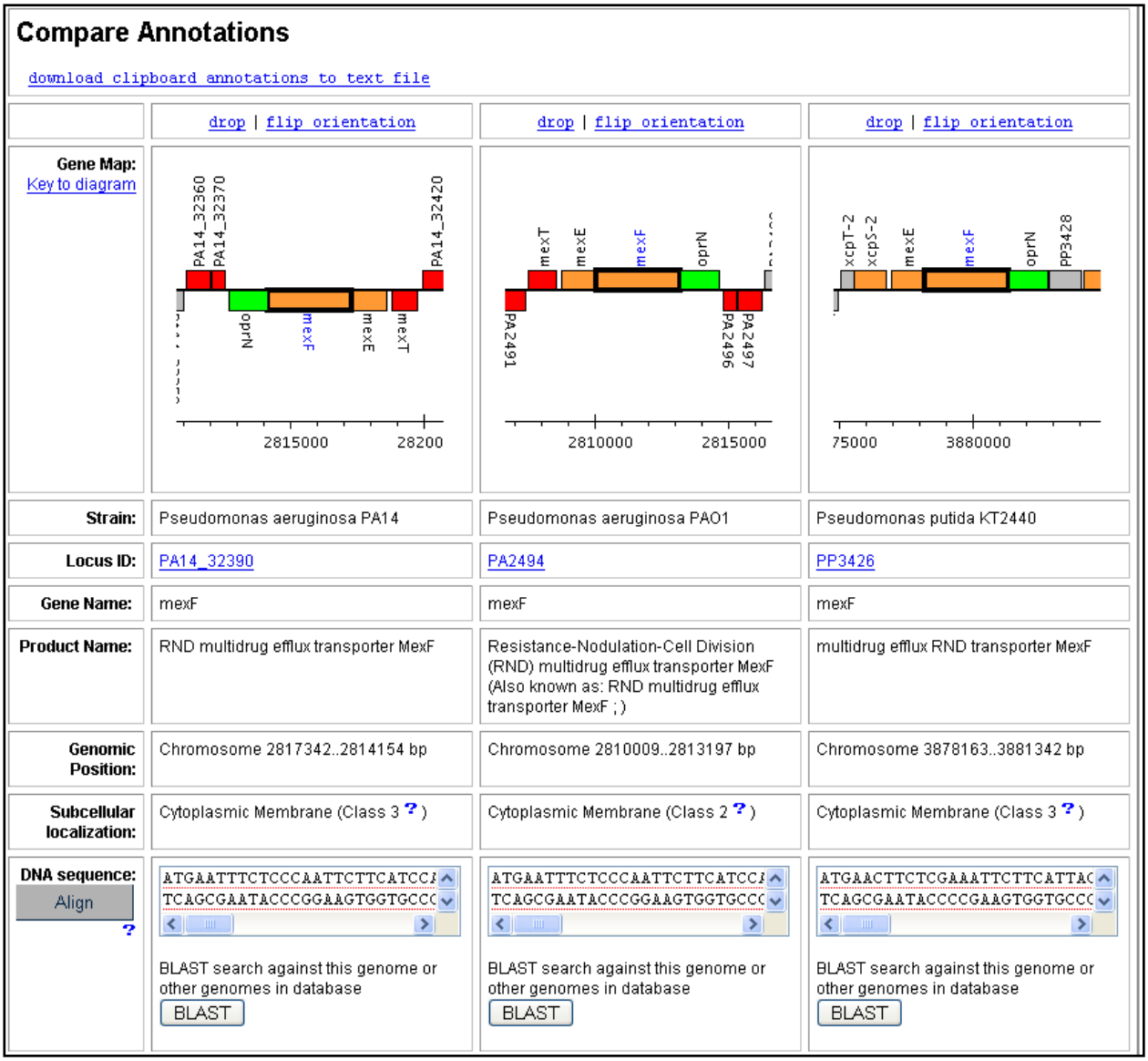
Species	Strain	Size of genome (Mbp)	Number of genes
Burkholderia cenocepacia	AU 1054	7.3	6632
Burkholderia cenocepacia	HI2424	7.6	7031
Burkholderia cepacia	AMMD	7.5	6724
****Burkholderia mallei***	ATCC 23344	5.8	5508
****Burkholderia pseudomallei***	K96243	7.3	5935

) are analyzed before they are added to the database. Some analyses include the prediction of subcellular localization using PSORTb v2.0 (Gardy, et al., 2005), prediction of protein classifications using Clusters of Orthologous Groups (COGs) (Tatusov, Galperin, Natale, & Koonin, 2000), and the prediction of protein domains using protein-domain families (PFAM) (Bateman, et al., 2002),. Without going into much detail about these predictions, it is important to note that there is a considerable amount of data stored in this database.

Burkholderia sp.	383	8.7	7826
****Burkholderia thailandensis***	E264	6.7	5714
****Burkholderia vietnamiensis***	G4	8.4	7863
****Burkholderia xenovorans***	LB400	9.8	9044

Table 1. Summary of species in *Burkholderia* Genome Database.
*** new species added to database**

Some of the main features of this database include a search interface for all genes and a variety of multi-genome comparison tools. A clipboard page allows the user to easily compare properties and functional predictions of added genes on a single page (see Figure 4). The *Pseudomonas* database also includes an update interface in which researchers can propose changes in annotation features to existing genes, or submit new annotations for genes they have discovered. An updates log provides a list of all changes made and names of researchers performing the changes. Furthermore, the website provides access to download entire genome annotations, protein sequences, DNA sequences, intergenic regions (the region between genes), or alternatively the user can specify the start and stop positions of the sequences they would like to download.



Compare Annotations

download clipboard annotations to text file

	drop | flip orientation	drop | flip orientation	drop | flip orientation
Gene Map: Key to diagram			
Strain:	Pseudomonas aeruginosa PA14	Pseudomonas aeruginosa PAO1	Pseudomonas putida KT2440
Locus ID:	PA14_32390	PA2494	PP3426
Gene Name:	mexF	mexF	mexF
Product Name:	RND multidrug efflux transporter MexF	Resistance-Nodulation-Cell Division (RND) multidrug efflux transporter MexF (Also known as: RND multidrug efflux transporter MexF ;)	multidrug efflux RND transporter MexF
Genomic Position:	Chromosome 2817342..2814154 bp	Chromosome 2810009..2813197 bp	Chromosome 3878163..3881342 bp
Subcellular localization:	Cytoplasmic Membrane (Class 3 ?)	Cytoplasmic Membrane (Class 2 ?)	Cytoplasmic Membrane (Class 3 ?)
DNA sequence: Align ?	ATGAATTTCTCCCAATTCTTCATCC… TCAGCGAATACCCGGAAGTGGTGCCC… BLAST search against this genome or other genomes in database BLAST	ATGAATTTCTCCCAATTCTTCATCC… TCAGCGAATACCCGGAAGTGGTGCCC… BLAST search against this genome or other genomes in database BLAST	ATGAACTTCTCGAAATTCTTCATTAC… TCAGCGAATACCCCGAAGTGGTGCCC… BLAST search against this genome or other genomes in database BLAST

This database also incorporates many tools developed by other researchers around the world so that they are accessible for researchers studying *Pseudomonas* species. Fist of all, there is a Basic Local Alignment and Search Tool (BLAST) interface to find regions of local similarity between sequences based on nucleotide or protein sequences (Altschul, Gish, Miller, Myers, & Lipman, 1990). Multiple sequence alignments can be performed using ClustalW to identify conserved regions of the DNA and provide information on generating phylogenetic/evolutionary trees (Chenna, et al., 2003).

A visualization tool used in this database is the Generic Genome Browser (GBrowse) in which tracks of genomic annotations and other information are displayed (Stein, et al., 2002). **Error! Reference source not found.** shows an example of the first few genes in *Pseudomonas aeruginosa* PA14 genome and a few additional tracks such as subcellular localization predictions, protein families (PFAM) predictions, Clusters of Ortholgous Groups (COG) predictions, and the predicted probabilities of genes being part of operons. This viewer also allows easy scrolling and zooming through the entire genome.

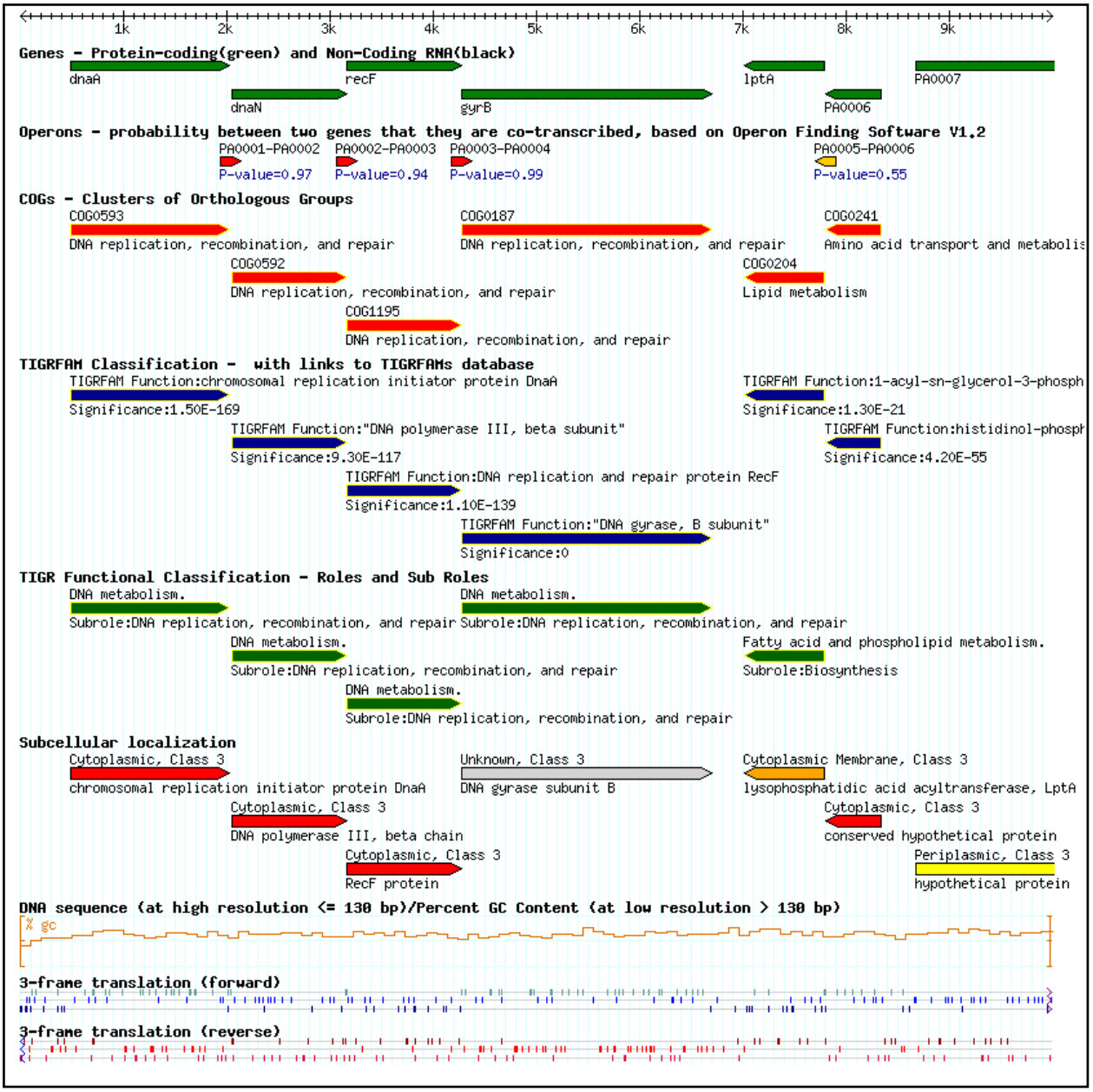


Figure 5. Generic Genome Browser view of annotation data.

Finally, PseudoCyc is another feature included in the *Pseudomonas* Genome Database. PseudoCyc is a tool based on EcoCyc and MetaCyc (Karp, Riley, Saier, Paulsen, Paley, & Pellegrini-Toole, 2000), which uses Pathway Tools (Karp, Riley, Saier, Paulsen, Paley, & Pellegrini-Toole, 2000) to provide an overview of all enzymatic reactions and metabolic pathways encoded by the genes in the genome of *Pseudomonas aeruginosa* PAO1.

3.2 The Burkholderia Genome Database

The *Burkholderia* Genome Database was set up to mimic the features of the original *Pseudomonas* Genome Database. All web pages were copied and modified to include information about *Burkholderia*, and the same database schema was used. In order to load the same amount of information for *Burkholderia* species into the new database, all of the same analyses were performed on each genome (subcellular localization, COG, PFAM, etc.). This was done using Perl scripts and all jobs were run on 'Buster the cluster' of 60 computer nodes. Once the analyses were complete, all data was loaded into the main *Burkholderia* database.

At the time of its first release, the *Burkholderia* Genome Database included data for 4 *Burkholderia* species (see Table 1). The website included the same search interface, BLAST page, sequence alignment tools, GBrowse viewer, downloadable sequences and links to other resources. There was no update interface for researchers since these genome annotations are not currently being updated. Also, a PseudoCyc-like interface was not set up using the Pathway-Tools software so users are unable to view metabolic pathways and reactions included in the genomes.

But overall, this database still provides a substantial amount of data for *Burkholderia* species so that researchers involved in studying the biological processes of these organisms can gain insight from their genomic properties. It is a valuable website designed to help researchers find new vaccines and treatments that will aid in preventing fatal effects in cystic fibrosis patients.

4 Improvements to the Databases

For the latter portion of my 8-month work term, I focused my attention on expanding the capabilities of the *Pseudomonas* and *Burkholderia* genome databases. First of all, I included 5 more genomes for each database. A summary of all of the species and their genome sizes is shown in

Species	Strain	Size of genome (Mbp)	Number of genes
Pseudomonas aeruginosa	PAO1	6.3	5570
Pseudomonas aeruginosa	PA14	6.5	5977
****Pseudomonas aeruginosa***	PA7	6.6	6369
Pseudomonas entomophila	L48	5.9	5275
Pseudomonas fluorescens	Pf-5	7.1	6233
****Pseudomonas fluorescens***	PfO-1	6.4	5833
****Pseudomonas mendocina***	ymp	5.1	4704
****Pseudomonas putida***	F1	6.0	5405
Pseudomonas putida	KT2440	6.2	5516
****Pseudomonas stutzeri***	A1501	4.6	4210
Pseudomonas syringae	B728a	6.1	5220
Pseudomonas syringae	DC3000	6.5	5849
Pseudomonas syringae phaseolicola	1448A	6.1	5436

Table 2. Summary of species in *Pseudomonas* Genome Database.
* new species added to database

Species	Strain	Size of genome (Mbp)	Number of genes
Burkholderia cenocepacia	AU 1054	7.3	6632
Burkholderia cenocepacia	HI2424	7.6	7031
Burkholderia cepacia	AMMD	7.5	6724
****Burkholderia mallei***	ATCC 23344	5.8	5508
****Burkholderia pseudomallei***	K96243	7.3	5935
Burkholderia sp.	383	8.7	7826
****Burkholderia thailandensis***	E264	6.7	5714
****Burkholderia vietnamiensis***	G4	8.4	7863
****Burkholderia xenovorans***	LB400	9.8	9044

Table 1. Summary of species in *Burkholderia* Genome Database.
*** new species added to database**

Species	Strain	Size of genome (Mbp)	Number of genes
Burkholderia cenocepacia	AU 1054	7.3	6632
Burkholderia cenocepacia	HI2424	7.6	7031
Burkholderia cepacia	AMMD	7.5	6724
****Burkholderia mallei***	ATCC 23344	5.8	5508
****Burkholderia pseudomallei***	K96243	7.3	5935
Burkholderia sp.	383	8.7	7826
****Burkholderia thailandensis***	E264	6.7	5714

and Table 1. These species also include strains that are non-pathogenic to humans for comparison to pathogenic strains to see differences and similarities. Some of these species are also studied in other fields of research. For example, *Pseudomonas fluorescens* is studied for its role in biological disease control and *Burkholderia vietnamiensis* is studied for its role in degrading pollutants. So these databases are not limited to researchers studying human pathogens.

***Burkholderia vietnamiensis**	G4	8.4	7863
***Burkholderia xenovorans**	LB400	9.8	9044

Table 1. Summary of species in *Burkholderia* Genome Database.
*** new species added to database**

Each genome was analyzed for subcellular localization, protein domains, protein function classifications, orthologs, and closest human homologs. Orthologs are genes in different species that have diverged by a speciation event but have similar function (Fulton, Li, Laird, Horsman, Roche, & Brinkman, 2006) so finding orthologous genes is important for prediction of gene function in newly sequenced genomes. Closest human homologs were found by using BLAST against the entire *Homo sapiens* protein sequence to make it easy to identify those genes that are closely related to human genes and may be difficult to use as vaccine targets.

In order to provide more information about orthologs, I developed a new stacked view of orthologs. This stacked view (see Figure 6) allows the user to view all orthologs of a given gene and provides the ability to easily compare their subcellular localization and the gene order around the gene of interest. Subcellular localization is often conserved among orthologs (Jensen, Ussery, & Brunak, 2003) so this may provide insight to predict the localization of those gene products assigned to an 'Unknown' localization by PSORTb. Figure 6 is a clear example of such a case.

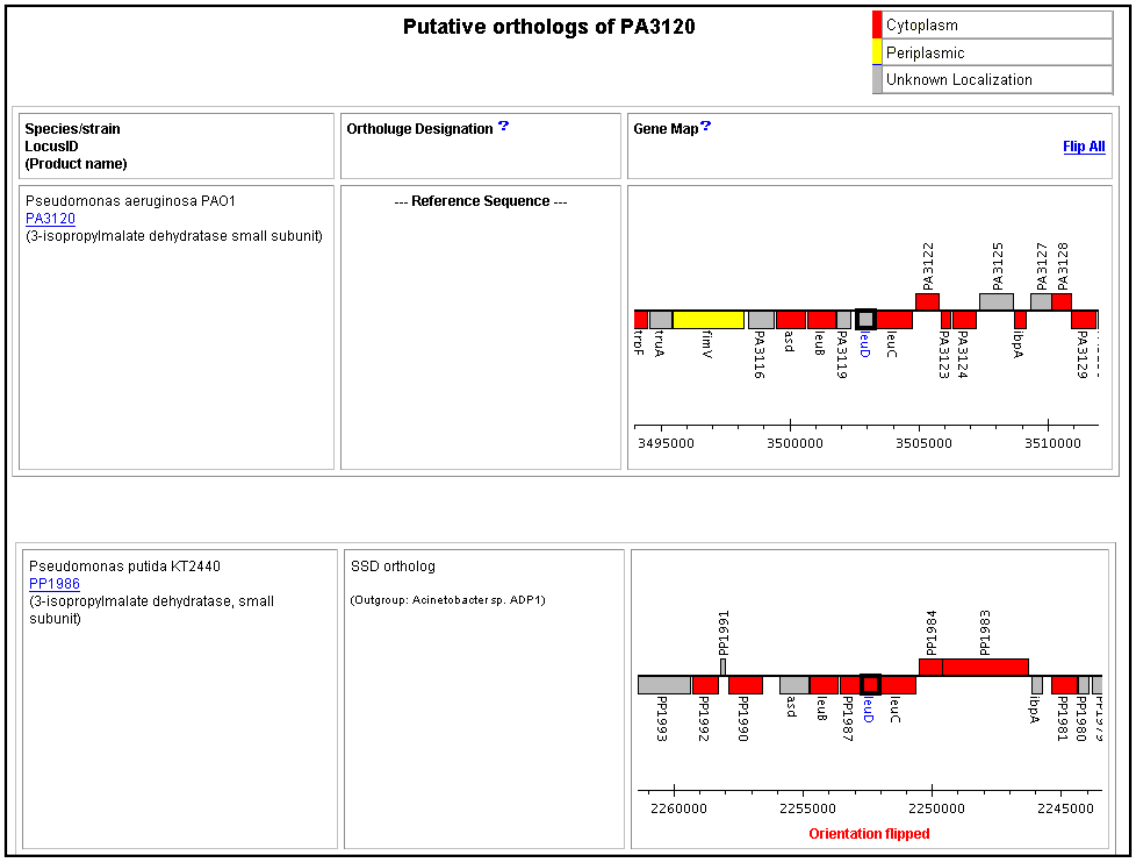


Figure 6. Stacked view of orthologs.

When considering protein domain predictions using PFAM, the database originally only stored the top hit, and that was considered the only domain present in the protein. However, proteins may have multiple domains (Bateman, et al., 2002) so it was necessary to make changes to the database to accommodate this. The database now stores all top hits above a specified cut-off score, with access to information about start and stop positions of the domains. All of this information was also loaded into the GBrowse viewer. As an example, Figure 7. Multiple protein domain predictions using PFAM. shows a gene (in green) with 5 different protein domain predictions (in blue), 4 of which

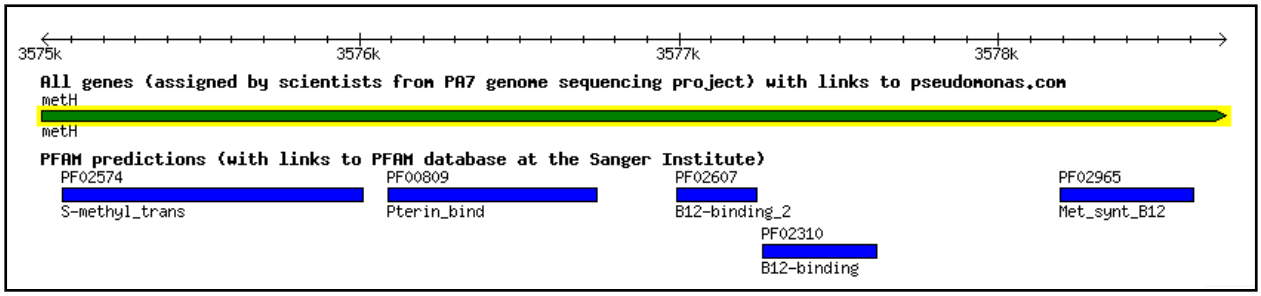


Figure 7. Multiple protein domain predictions using PFAM.

would have been missed in the original database.

Another important new feature that I developed was the combined search page, which allows the user to query both the *Burkholderia* and *Pseudomonas* databases. This was one of the main reasons both databases were hosted on a single machine. This feature is important since it allows researchers access to more comparative genomics methods. The search results limit was also increased to 8000 from 6000 to accommodate larger searches. Furthermore, filters were included for the user to filter the results based on the presence or absence of transposon mutant data and/or human homologs. These filters can also be used to query the databases without any other input, to get all genes without human homologs, for example. All of these changes to the search interface

provide stronger methods to analyze the genomic data in order to study the pathogenic properties of these organisms.

A few additional links to other available tools were also included in the updated version of the website. Firstly, links to the SYSTOMONAS database (Choi, et al., 2007) were included for access to information related to experimental data, and predictions of cellular processes and metabolic networks. In addition to this, links were also added to WebACT (Abbott, Aanensen, & Bentley, 2007), a tool that provides data for whole genome sequence alignments. It uses the Artemis Comparison Tool (ACT) (Carver, Rutherford, Berriman, Rajandream, Barrell, & Parkhill, 2005) to visualize evolutionary events by comparing entire genomes to each other. This provides insight on the evolution of the species being compared and which genes may have provided some strains with more pathogenic capabilities.

5 Conclusions

Overall, through this 8-month work term, I have reached 2 main objectives: the release of the *Burkholderia* Genome Database and the development of new, insightful research tools that may prove to be very useful in studying the pathogenicity of *Pseudomonas* and *Burkholderia* species at a higher level. In turn, I hope that this will be a stepping stone so that in the near future researchers are able to provide improved treatment methods and preventative vaccines to cystic fibrosis patients that may be affected by these bacteria during their lifetime.

6 Recommendations for Future Work

Just as with any scientific project, the *Pseudomonas* and *Burkholderia* Genome Databases can still be improved! First of all, it is absolutely imperative to write scripts that will check the status of both databases so that if there are any problems, they can be addressed in a quick and painless fashion. These databases are used widely around the world, so any interruptions in access to the websites may be an inconvenience to researchers.

Another important analysis to perform again would be Ortholuge since there were many new species added to each database. This analysis has already been performed between all species previously in the database, but needs to be completed for all species against each other to provide an inclusive set of orthologs between all species.

In addition to this, it will be important to include information about genes found in intergenic regions in the genomes. A student in the Brinkman lab is investigating genes in intergenic regions in *Pseudomonas aeruginosa* PAO1, so it would be ideal to use this method to identify other genes in all other genomes and include the information in the database.

Signature-tagged mutagenesis data could be included so that researchers can overlay the data with genomic information already in the database. Signature-tagged mutagenesis is a type of experiment that introduces mutations in the genome and the resulting increase or decrease in pathogenicity or death of the mutants is used to identify genes that may be involved in the infection mode of the organism. This information could be valuable in experiments as more and more data is produced and included in the database. Furthermore, additional experiments could be included,

such as enzyme assays, so that researchers have access to all of this information from a single website.

Some of the main long term goals for this project are aimed at providing improved comparative analyses through an advanced search page that allows users to perform a variety of different searches. For example, one may want the option to select all genes that are unique to a specific strain compared to a set of other strains. These types of comparative search queries would greatly enhance the ability of these databases to provide information that researchers are looking for.

It will also be very valuable to integrate further tools for whole genome alignments. In addition to the Artemis Comparison Tool, Mauve (Darling, Mau, Blattner, & Perna, 2004) is an alternative method for producing these alignments. Mauve is geared more towards large-scale analysis of the genomes and evolutionary events leading to rearrangements and other changes of the DNA. Incorporating both tools will allow researchers to use either method as suitable to their needs.

Finally, it will also be important to develop a protein-protein interaction network to link molecules that may be involved in signaling pathways and reactions. Also, integration of the Cytoscape (Shannon, et al., 2003) project will be useful in visualization of these interaction networks. This portion of the website may become crucial in understanding the role of certain proteins in pathways that allow these bacterial species to infect humans and evade the host immune system.

Overall, much has been done to improve these databases since their first release, and in the future they will continue to advance as more research is done in bioinformatics and more tools become widely available and commonly utilized. Our goal is that this database will be used as a model for other scientists to study genomes of other bacteria so that researchers can come up with solutions and strategies in this fight against bacterial pathogens.

References

Abbott, J., Aanensen, D., & Bentley, S. (2007). WebACT: An Online Genome Comparison Suite. *Methods of Molecular Biology , 395*, 57-74.

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology , 215* (3), 403-10.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S., et al. (2002). The Pfam protein families database. *Nucleic Acids Research , 30* (1), 276-80.

Canadian Cystic Fibrosis Foundation. (2007, March 13). *Canadian Cystic Fibrosis Foundation*. Retrieved 11 29, 2007, from CCFF: about cystic fibrosis: disease information: http://www.cysticfibrosis.ca/page.asp?id=1

Carver, T., Rutherford, K., Berriman, M., Rajandream, M., Barrell, B., & Parkhill, J. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics , 21* (16), 3422-3.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T., Higgins, D., et al. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research , 31* (13), 3497-500.

Choi, C., Munch, R., Leupold, S., Klein, J., Seigel, I., Thielen, B., et al. (2007). SYSTOMONAS -- an integrated database for systems biology analysis of Pseudomonas. *Nucleic Acids Research , 35* (Database issue), D533-7.

Darling, A., Mau, B., Blattner, F., & Perna, N. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research , 14* (7), 1394-403.

Fulton, D., Li, Y., Laird, M., Horsman, B., Roche, F., & Brinkman, F. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics , 7*, 270.

Gardy, J., Laird, M., Chen, F., Rey, S., Walsh, C., Ester, M., et al. (2005). PSORTb c2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics , 21* (5), 617-23.

Jensen, L., Ussery, D., & Brunak, S. (2003). Functionality of system components: conservation of protein function in protein feature space. *Genome Research , 13* (11), 2444-9.

Karp, P., Riley, M., Saier, M., Paulsen, I., Paley, S., & Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Research , 28* (1), 56-9.

Shannon, P., Markiel, A., Ozier, O., Balaga, N., Wang, J., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research , 13* (11), 2498-504.

Stein, L., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., et al. (2002). The generic genome brower: a building block for a model organism system database. *Genome Research , 10*, 1599-610.

Tatusov, R., Galperin, M., Natale, D., & Koonin, E. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research , 28* (1), 33-6.

Winsor GL, L. R. (2005). Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res , 33* (Database issue), D338-43.