A Computational and Biological Analysis of the Proteins in Acinetobacter baumannii

Triple A: Adam Klie, Alexander Lui and Angela Frentzen
adamklie13@gmail.com, alex.lui858@gmail.com and angelafrentzen@gmail.com

Introduction:

Acinetobacter baumannii is a pathogenic bacteria that is causing serious health concerns in hospitals around the world. A member of bacterial group *Acinetobacter* commonly found in soil and water, *A. baumannii* is a particularly pathogenic strain responsible for about 80% *Acinetobacter* infections [1]. The gram-negative strain is very common in hospitals due to its has an innate ability to colonize and survive on artificial surfaces as well as its "opportunistic" behavior [2]. This organism is not typically harmful to healthy individuals, but commonly targets those most vulnerable to infection in healthcare settings [1].

A. baumannii has grown to be quite adept at surviving in hospitals. It has shown the ability to survive desiccation and disinfection and can grow on dry surfaces due to its ability to form biofilms [3] and also acquired multidrug resistance [4]. The bacterium has shown an ability to survive almost all available antibiotics, rendering an infection very difficulty to treat. The multidrug resistance of this pathogen, coupled with its ability to live and grow in many different environments within hospitals, make infection from this strain of bacteria a serious health concern to doctors everywhere.

In an attempt to better understand the basic mechanisms by which this bacteria can survive and grow as well as the factors that make it virulent, we analyzed partially annotated protein sequences and annotated them for function. Specifically, we used publicly available sequence analysis tools and databases in an attempt to assign function to the proteome of this organism and detect the presence of any proteins that may help us understand how this bacteria has become such an issue in hospitals.

Methods:

Our input for the overall analysis was the partially annotated proteome of *A. baumannii* in fasta format. We chose a subset of sequences to analyze and annotate and ran them through the following computational pipeline and subsequent biological analysis. The following tools and databases were all queried and parsed using scripts were called in a singular main function that ran the entire python pipeline as a single script. In totality, we computational annotated 1,050 sequences (#450-549, 750-849, 900-999, 1000-1749) using this method.

BLAST - In order to assign function to our sequences we began by using Protein BLAST (pblast) [5]. In order to speed-up each alignment and not run into any issues querying the BLAST servers, we downloaded the pblast tool and ran our sequences against the entire Uniprot database [6] locally. To streamline the process, we ran all the sequences as one query, generating a single raw data file from the BLAST run. We then needed to parse the raw data file for functional keywords for each hit on each of

our sample proteins. To do so, we utilized the Biopython BLAST NCBIXML module that allowed us to parse the raw xml file (we also saved a plain text file) for the accession number and protein name of all our protein hits in the database. In doing this, we only saved the accession numbers of the top hit for each query sequence, as long as it had an E-value less than 1. The accession number was further utilized to query the Uniprot online database to pull keywords associated with the database proteins. The Biopython ExPASy and SwissProt modules were used for this purpose. The protein title from BLAST was concatenated to the keywords pulled from Uniprot, separated by semicolons and output the second column of a tsv file (the first column of the output tsv corresponds to the accession number of the query protein).

PFAM - To both supplement are BLAST hits and provide function for those proteins without good BLAST annotations, we used the locally downloadable tool HMMER [7] and another python script to query the Pfam database [8]. Specifically, we used hmmscan to query our sequences against the Pfam hidden markov model (HMM). We again ran all of our sequences in a single run against the HMM and generated tabular form output which we saved. To parse this output file for functional keywords, we again relied on another Biopython module, this time utilizing the SearchIO module. Using this module, we pulled all the description fields for each hit with an E-value of less than 1, separated them by semicolon and output them to the third column of our final output tsv

PROSITE - Querying the entire Prosite regular expression motif database was very similar to the last two steps. We first downloaded pftools, [9] ExPASy's Prosite scanning tool. We also decided to use the option to exclude motifs with a high probability of occurrence. We queried all of our proteins in a single run and generated a raw text file that was relatively easy to parse using simple python text splitting functions. From this raw data file (which we also saved), we again pulled hit descriptions this time for every hit and output them in the fourth column in the same aforementioned format.

KEGG - In an attempt to understand the greater pathway functionality of each protein, we queried the KEGG database for terms [10]. To do so, we utilized the gene identification number provided in the partially annotated input file. The Bioservices.kegg KEGG module was used to query the KEGG database and pull terms in the form KEGG term [KEGG ID]. These terms were compiled for each protein in the same format as for the other terms and output to the fifth column of the tsv.

GENE ONTOLOGY (GO) -The process of assigning GO terms for our sequences involved the use of mapping files provided by the Gene Ontology consortium [11]. Through the use of many python dictionaries, we took the accession numbers from all the Uniprot, Pfam and Prosite hits we found and mapped them to their already annotated GO terms. This process was done by parsing these downloadable mapping files in separate python scripts and cross referencing them to the hit accession numbers we had already parsed and saved from our different output. The terms were again semicolon separated with each individual hit in the form GO term [GO: ID]. This output was then put in the 6 column.

Comments - We assigned functions to most of the first 100 proteins we selected (#450-549) based on the pipeline output. Since about 90% of the proteins had good BLAST hits, we looked at this field first. We then consulted the Pfam and Prosite results to either supplement the BLAST hit, or a find function if there was no good BLAST hit. Some annotations were difficult to understand because there were so many keywords. In this circumstance we looked at KEGG pathways and GO terms for more information. Each additional annotation often confirmed our analysis of the protein. When we found a particularly interesting function (e.g. multidrug resistance proteins), we checked the BLAST output file to verify the expected values and assess our confidence in the annotation. We manually commented the assigned 100 proteins, and for the extra 950, we did searches for functional keywords that we were particularly interested in.

Results:

Annotating 100 proteins with our pipeline took on average about 30 minutes. While manually commenting these results, we identified a few heavily enriched biological pathways. Metabolism was understandably the most common annotation and we also found many proteins involved in ribosomal synthesis and the process of translation. We also found less common but more interesting protein functions in our first 100 sequences. To supplement these findings, we were able to complete the computerized annotation of 1,050 sequences in the span of about two days using approximately 5-6 hours of total runtime. After annotating the 100's of hits we searched for specific molecular functions in our results that were likely to provide special insight into the bacterium and unique human applications.

Transposable Elements - Pipeline results showed several candidate proteins that may `have functions as transposable elements. Of these results, ABO12974, ABO13004, ABS90240 had listed associations with specific IS (insertion sequences). NCBI states that these families are grouped by genetic organization, transposition reactions, terminal domains and target site classifications [12]. Since IS families are already so well annotated, the BLAST hits with IS associations may allow for a more in depth understanding of the structure or mechanism involved in the bacteria's transposable DNA elements. Proteins ABO13445 and ABS90276 shared a common defining signature. Both had Pfam annotations of 'transposase zinc-ribbon domain'. This binding domain is associated with several transposase IS families, and is thought to allow DNA binding [13]. These extrapolations may help narrow down biological targets that can help analyze these proteins in a wet lab. Lastly, protein ABO12975 had blatantly obvious references to transposases in BLAST, Pfam and Gene Ontology results. This protein seems to be the most obvious transposase in the 1,050 proteome sequences analyzed.

Bacterial Virulence Factors - The computerized annotation of our sequences and the hand annotation in the comments field yielded many database hits that had keywords associated with bacterial virulence factors. We define virulence factors as those proteins that have function associated with making the bacteria a better pathogen and can include a variety of different types of functions and mechanisms.

Many such potential virulence factors identified in the input sequences were related to the bacteria's ability to respond to its environment. These included multiple hits with annotations related to immunoglobulin binding and downregulation of the host immune system (ABO13854, ABO10924). These proteins had good BLAST hits with subsequent GO terms that indicated that they had the potential to bind host cell antibodies and in some fashion and reduce the host's immune response. We also were able to find a multitude of proteins with annotations associated to zeta and AbiEii toxins. Zeta proteins have known functions in killing plasmid-free bacteria and is bactericidal in nature [14]. AbiEii toxins have known functions in preventing infection from bacteriophages and could further boost the ability of the *A. baumannii* survive and outcompete other strains [15]. Other noted annotations that may have functions in helping this bacteria navigate and survive its environment were sporulation related proteins (ABO13822). These proteins may be functional in helping the cell become dormant and allowing it to survive in the face of environmental stresses, including the presence of bactericides, antibiotics and desiccated surfaces.

Capsule biosynthesis proteins were also noted in our annotations (ABO13718). Bacterial capsules have been shown to make strains more virulent by way of providing a physical barrier against antibodies and antibiotics. Furthermore, biofilm production related proteins found in the annotations (ABO11966, ABO11371) and may play a role in bacteria's ability to survive on artificial surfaces for extended periods. Finally, a few of our most interesting hits were related the bacteria's ability to adhere to epithelial cells of a host. We found multiple annotated sequences that were related to outer membrane proteins, which are known to play roles in adherence in other bacteria [16]. We also found a few very interesting hits related to the pathogenic bacterium that causes cholera, *Vibrio cholerae*. Specifically, we found queries (ABO13283, ABO11730) with cholera toxin related domain annotations in Pfam. These domains both are tied to the adherence of the cell to the host epithelium, something *V. cholerae* is expert at.

Antibiotics Resistance - The data showed proteins that are associated with several mechanisms important to antibiotics resistance. Proteins ABO12625, ABO13835 and ABO13834 are shown to have AcrB, AcrD and AcrF efflux pump annotations that contribute to antibiotic resistance by pumping out invading toxins [17]. These membrane proteins are part of the ABC (ATP binding cassette) superfamily. Analogously, major facilitator superfamily (MFS) profiles which are fairly similar to those of the ABC superfamily, have also exhibited signs of antibiotic resistance [18]. We noticed an interesting protein, ABO10644 that had both an MFS profile and virulence promoting factors. These two put together could help improve the virulent potency of this bacteria, capable of surviving antibiotics and infecting other cells. One notable protein, ABO12625 had efflux pumps, MFS profiles, transmembrane transport, and an "antibiotic resistance" hit, making it a very attractive candidate for further wet and dry lab studies

Discussion:

In investigating our subset of the *A. baumannii* proteome, we generated annotations on many sequences that had very interesting functions pertinent to the bacteria's virulence and effectiveness in hospital

settings. These annotated proteins included those which had function in allowing the bacteria to survive under different environmental stresses and conditions that other organisms are unable to. Specifically, we found multiple sporulation and dormancy related proteins that could potentially allow the bacteria to survive periods of desiccation or intense heating or cooling. Along those same lines, we annotated a few proteins that may aid the bacteria in the production of biofilms, further conferring the organism with the ability to survive on many different hospital surfaces and aiding the transmission of the infection from person to person. Furthermore, we identified two toxins within the bacterium that seem to be related to toxins produced by *V. cholerae* and have been previously known to allow the cell to adhere to epithelial tissue in a host. Outer membrane proteins, previously shown to also aid in cell adherence to epithelium and other surfaces, were also present in our set of proteins. The combination of these different proteins are most likely part of what makes this organism so adept at surviving and then subsequently colonizing a host in a hospital setting.

The ability of bacteria to take up DNA from the environment is a process that largely controls bacteria's fantastic ability to adapt and survive. The ability to take up external genes would be largely ineffective to improve survival if bacteria did not have the ability to incorporate these new genes into their genomes. Transposable elements make this possible by coding for transposases that are involved in cutting or copying genomic segments and inserting them into a new place in the genome. This process allows for the appropriate integration of extracellular genomic DNA into the organism's genome to improve survival. Bennett emphasizes the importance of transposable elements in, "indicating that genomes of all bacteria can be a single global gene pool into which most, if not all bacteria can dip for genes necessary for survival" [19]. This realization makes the study of transposable elements in bacteria important, because it has great implications for allowing integration of desirable genes such as antibiotic resistance or other virulence factors into the genome.

We also noted many different kinds of protein sequences that had the common function of fighting off bactericidal elements such as antibiotics, competing organisms, viral infection and the host immune system. Of most interest where the multiple efflux pump proteins annotated in our sequences that most likely confer the organism with the ability to pump multiple drugs out of the cell. After taking BICD110 (cellular biology) with Dr. Forbes, we learned that multidrug resistance in cancer cells came about after one or two treatments with many different drug cocktails. It is quite possible that a similar drug resistance story may be occurring here. According to Norbert Kartner and Victor Ling, a certain membrane protein "P-glycoprotein is associated with pumping these drugs out of cancer cells giving these immunity [20]. Given that these efflux pumps have been previously observed in *A. baumannii* [21] it makes sense that we were able to annotate a few in our sequences. Further investigation also showed many other proteins that could be involved in the breakdown or inactivate the effectiveness of different drugs (see supplementary data). Given the ever-present nature of antibiotics in hospital settings, it makes sense to see a bacteria such as this develop antibiotic resistance.

In taking a step back and looking at all the aforementioned functional annotations as a whole, it is apparent that this bacteria has adapted to become amazingly adept at surviving in hosptitals. It has a

multitude of defenses against common bacteria killing agents commonly found in hospitals and even other areas. It makes sense that the selective pressures dictated by the hospital environment (i.e. constant stream of antibiotics, compromised immunity, dry and sanitized surfaces) that certain bacteria would evolve ways to survive and thrive in said environment. *A. baumannii* appears to be that bacteria and will continue to pose an issue in hospital settings around the world without the development of new methods to fight the infection and treat patients and healthcare settings.

Challenges - About 10% of our results did not have a BLAST hit, which necessitated pulling functional associations from Pfam, Prosite, KEGG pathways and GO terms. We also had issues with understanding the level of significance of our hits from just our output tsv file. When proteins with an interesting BLAST hit were annotated, it became time consuming to manually cross reference the gene ID's to the BLAST output file and verify the expected values to infer confidence for the annotation. There were additional difficulties in obtaining KEGG and Prosite annotations. These two databases often failed to produce hits, giving little to no information to obtain a functional inference. However it was equally as challenging to address the opposite problem of having too many annotations with several diverse functional hits.

Improvements - Several simple improvements could be made to our pipeline to improve its robustness and ease of use. To begin with, we did not display the expected values outputted from blast in our final table. As mentioned above, this made commenting the results more difficult because the confidence levels of a hit could not be easily accessed directly from the final table. This simple change would result in increased efficiency in assigning function. Another easy optimization that can be added to the pipeline is altering the KEGG pathway queries. The data contained very few KEGG pathway hits because KEGG was queried using the original bacteria's protein IDs rather than the IDs of the BLAST hits, which were much more likely to be annotated. Changing this query in our pipeline would result in more information that will allow for further protein assessment. These simple scripting changes would greatly optimize our ability to annotate protein function from the pipeline results.

When selecting BLAST hits, the current pipeline takes the singular top hit from the results with an expected value less than one. Taking a single hit is a somewhat arbitrary cutoff that could be optimized by further taking into account the orthology of the organism in the process, since organisms with similar orthology tend to overlap in function. This means that in filtering blast, just taking the top hit might not be the best option. A hit with a lower score will likely give more accurate functional information if the protein is homologous. We can optimize this process by using one of many widely available tools. Reciprocal BLAST infers orthology by doing a blast search on organism A's protein A, taking the top hit that contains organism B and protein B, and blasting protein B against organism A to see if this reciprocal search also yields protein A as the top hit [22]. If protein A is listed as a top hit, orthology can be inferred. The list of orthologs between two genomes is stored in databases such as Homologene and InParanoid [22]. However our little-known bacteria may not have data in places, so using reciprocal BLAST may be a useful tool to infer orthology. This can be applied to our pipeline by running BLAST hits against an ortholog database before choosing which hit to take into account for further annotation.

Conclusion:

The value of database tools like BLAST, Pfam, Prosite and KEGG are readily apparent from this project. These tools made it possible to take a fasta file of a largely unannotated bacterial genome and infer protein functions that may have huge implications for the future of proteome research. From this short project we've inferred that the proteins of *Acinetobacter baumannii* play a role in in genetic transposition, organism toxicity and pathogenesis, and bacterial resistance. This information can further be used in wet labs to verify protein function and alter them to help humanity's antibiotic resistance epidemic and several other issues.

References:

[1] - "Healthcare-associated Infections." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, 24 Nov. 2010. Web. 11 June 2017.
[2] - Howard, Aoife et al. "Acinetobacter Baumannii: An Emerging Opportunistic Pathogen." Virulence 3.3 (2012): 243–250. PMC. Web. 11 June 2017.
[3] - Espinal, P., S. MartÃ, and J. Vila. "Effect of Biofilm Formation on the Survival of Acinetobacter Baumannii on Dry Surfaces." Journal of Hospital Infection 80.1 (2012): 56-60. Web.
[4] - Dijkshoorn, Lenie, Alexandr Nemec, and Harald Seifert. "An Increasing Threat in Hospitals: Multidrug-resistant Acinetobacter Baumannii." Nature Reviews Microbiology 5.12 (2007): 939-51. Web.
[5] - Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed
[6] - The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45: D158-D169 (2017)
[7] - Accelerated profile HMM searches. S. R. Eddy. PLoS Comp. Biol., 7:e1002195, 2011.
[8] - Finn, Robert D. et al. "Pfam: The Protein Families Database." Nucleic Acids Research 42.Database issue (2014): D222–D230. PMC. Web. 11 June 2017.
[9] - Sigrist, Christian J. A. et al. "PROSITE, a Protein Domain Database for Functional Characterization and Annotation." Nucleic Acids Research 38.Database issue (2010): D161–D166. PMC. Web. 11 June 2017.
[10] - Kanehisa, Minoru, and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Research 28.1 (2000): 27–30. Print.
[11] - The Gene Ontology Consortium. Gene Ontology Consortium: going forward. (2015) Nucl Acids Res 43 Database issue D1049–D1056. Online at Nucleic Acids Research.
[12] - Mahillon, Jacques, and Michael Chandler. "Insertion Sequences." Microbiology and Molecular Biology Reviews 62.3 (1998): 725–774. Print.
[13] - EMBL-EBI, InterPro. "InterPro." Transposase, Zinc-ribbon (IPR024442). N.p., n.d. Web. 11 June 2017.

[14] - Zielenkiewicz, Urszula, and Piotr Cegłowski. “The Toxin-Antitoxin System of the Streptococcal Plasmid pSM19035.” Journal of Bacteriology 187.17 (2005): 6094–6105. PMC. Web. 11 June 2017.
[15] - Dy, Ron L. et al. “A Widespread Bacteriophage Abortive Infection System Functions through a Type IV Toxin–antitoxin Mechanism.” Nucleic Acids Research 42.7 (2014): 4590–4605. PMC. Web. 11 June 2017.
[16] - Koebnik, R., Locher, K. P. and Van Gelder, P. (2000), Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Molecular Microbiology, 37: 239–253. doi:10.1046/j.1365-2958.2000.01983.x
[17] “Crystal structure of bacterial multidrug efflux transporter AcrB.” Murakami S, Nakashima R, Yamashita E, Yamaguchi A.
[18] “Modulation of Bacterial Multidrug Resistance Efflux Pumps of the Major Facilitator Superfamily (MFS).” Sanath Kumar, Mun Mun Mukherjee, and Manuel F. Varela.
[19] - Bennett, P M. “Plasmid Encoded Antibiotic Resistance: Acquisition and Transfer of Antibiotic Resistance Genes in Bacteria.” British Journal of Pharmacology 153.Suppl 1 (2008): S347–S357. PMC. Web. 10 June 2017.
[20] “Multidrug Resistance in Cancer.” Norbert Kartner and Victor Ling.
[21] - Magnet, Sophie, Patrice Courvalin, and Thierry Lambert. “Resistance-Nodulation-Cell Division-Type Efflux Pump Involved in Aminoglycoside Resistance in Acinetobacter Baumannii Strain BM4454.” Antimicrobial Agents and Chemotherapy 45.12 (2001): 3375–3380. PMC. Web. 11 June 2017.
[22] - "Determining Orthology." Orthology. N.p., n.d. Web. 11 June 2017.