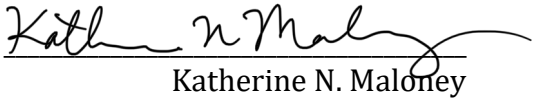
Exploring Biosynthetic Gene Clusters In Amicoumacin Producing Organisms

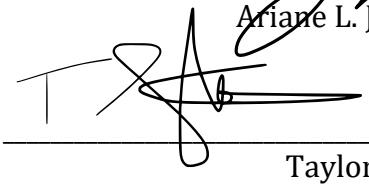
by

Amanda Koontz

Submitted in partial fulfillment of the requirements for graduation as an Honors Scholar at Point Loma Nazarene University, San Diego, California on May 7, 2022.

Approved by 
Katherine N. Maloney


Ariane L. Jansma


Taylor Steele

Date: 4/3/2023

Exploring Biosynthetic Gene Clusters in Amicoumacin Producing Organisms

Amanda Koontz and Katherine Maloney
Department of Chemistry, Point Loma Nazarene University, San Diego, CA

Abstract

Natural products are small organic molecules which are presumed to increase the overall fitness of an organism. Their synthesis is encoded by biosynthetic gene clusters wherein analogous sequences can produce similar chemical products. This project utilizes computational tools to investigate enzymes associated with the amicoumacin-producing gene cluster of CB729.

Introduction

Background

The central dogma of molecular biology describes the information flow from DNA to RNA to proteins. These proteins, also known as enzymes, supply the perfect environment for chemical reactions to occur, eventually resulting in the production of specific molecules. Because of the relationship between genetic material and the production of molecules, there is a vast amount of information that can be used to learn more about how different organisms may produce the same or different chemical products. The information encoded in the DNA directly leads to the production of a particular molecule and by using computational tools to examine the DNA sequence, we can often predict the structure of a molecule.

While some biosynthetic pathways result in the formation of molecules necessary to survival, other pathways result in the production of natural products. Natural products are small, organic molecules, with no recognizable role in the normal growth and development of an organism, although they are presumed to increase the overall fitness of the producer or organisms that share its immediate environment. Genes that lead to natural product biosynthesis are often found near to each other on the DNA in a so-called biosynthetic gene cluster, or BGC. Once characterized, the presence of a particular BGC can be used to identify other organisms with the potential to produce the same or a similar natural product. In tandem with natural product isolation, the characterization of BGCs and biosynthetic pathways are an interesting area of research. Slight variations in the DNA sequence may result in the production of different enzymes and the production of slightly different natural products.

Citrus Greening Disease, Citrus Bacteria 729, and the Amicoumacin Family

One common reason organisms make natural products is to kill competitors, as is the case with antibiotics. In the US, Citrus Greening Disease, caused by the bacterium *Candidatus* Liberibacter asiaticus (*C*Las), has devastated citrus production by more than 70%.[1] In the search

for antibiotic natural products, citrus bacteria (CB) living within the microbiome of Florida citrus trees was screened for activity against *Liberibacter crescens*, which is a culturable surrogate for *C*Las.[2] The intention of this screening was to isolate bacterial strains that were producing antibiotic natural products to treat or prevent Citrus Greening Disease. Of the 1656 bacterial isolates found from Florida trees, 37 showed bacterial inhibition.[3] By using normal phase flash chromatography and HPLC, as well as NMR and LCMS data, CB729 was identified as a producer of a family of natural products called amicoumacins.

The amicoumacins are a family of dihydroisocoumarins which have been previously identified in other organisms as displaying antibacterial and antifungal activities.[3] Found in both Gram-positive *Bacillus* species and in the Gram-negative bacterium *X. bovienii*, amicoumacins are produced in a hybrid nonribosomal peptide synthase-polyketide synthetase (NRPS+PKS) pathway which results in active products.[3] This project focused on exploring the amicoumacin producing pathway in CB729 in comparison to other characterized biosynthetic pathways in various organisms.

Experimental Details

Natural Product Isolation

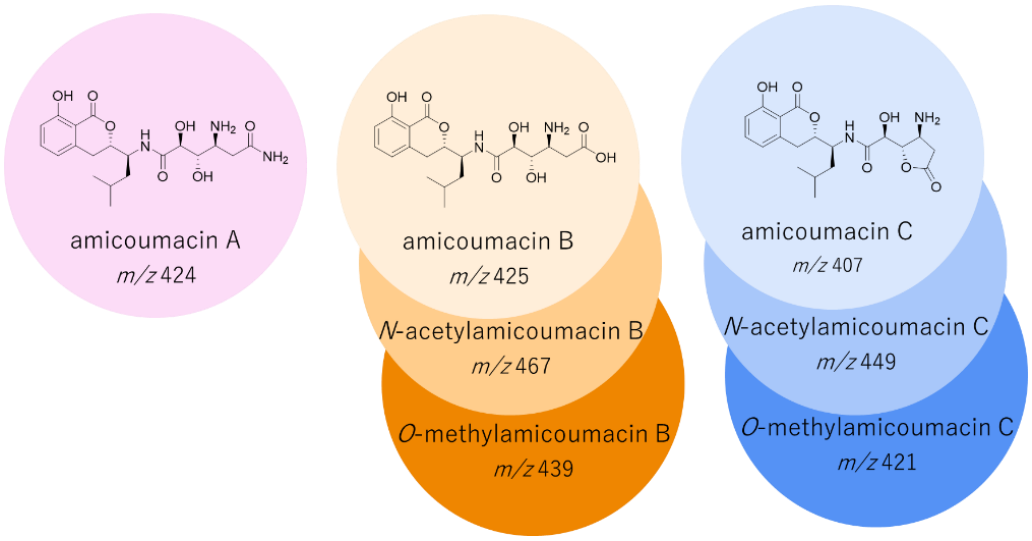


Figure 1. Amicoumacins identified in CB729 fractions by LCMS.

Within a culture of CB729, amicoumacins A, B, C, and *N*-acetyl and *O*-methyl derivatives of amicoumacins B and C were identified through fractionation and analysis (Figure

1). The major compound produced by this culture was *N*-acetylamicoumacin C, which is generally categorized as an inactive amicoumacin.

AntiSMASH

To identify the possible amicoumacin BGC, we used the antibiotics and secondary metabolites analysis shell, or antiSMASH, a pipeline for the mining of genetic data in the search for natural product BGCs.[4] It is a free, web-based platform written in Python which allows for the identification and analysis of BGCs in both fungal and bacterial genomes. The genome assembly for CB729 was provided by Alex Blacutt. By inputting the bacterial genome sequence for CB729, eleven regions were recognized as possible BGCs, some with a high similarity to known clusters and natural products. The only secondary metabolite region recognized as an NRPS+PKS type mechanism was Node 4.1, which has a 18% similarity to the zwittermicin A BGC produced by *Bacillus cereus*.

Clinker

Clinker is an open-source software used for generating gene cluster comparison figures.[5] When used in the command line, clinker takes GenBank files and aligns sequences based on genetic similarity. Figure 2 shows a BGC comparison made between CB729 Node 4.1, as recognized in antiSMASH, and *Bacillus subtilis* 1779, a known amicoumacin-producing marine bacterium.[6] *B. subtilis* 1779 was identified through a literature search of amicoumacin-producing organisms with characterized biosynthetic pathways. Alignment of the gene clusters show a significant similarity between sequences, which in theory would result in a similar composition of natural products produced by each organism. A majority of the genes associated with amicoumacin production in *B. subtilis* are aligned with genes from CB729 Node 4.1 and are labeled in Figure 2.

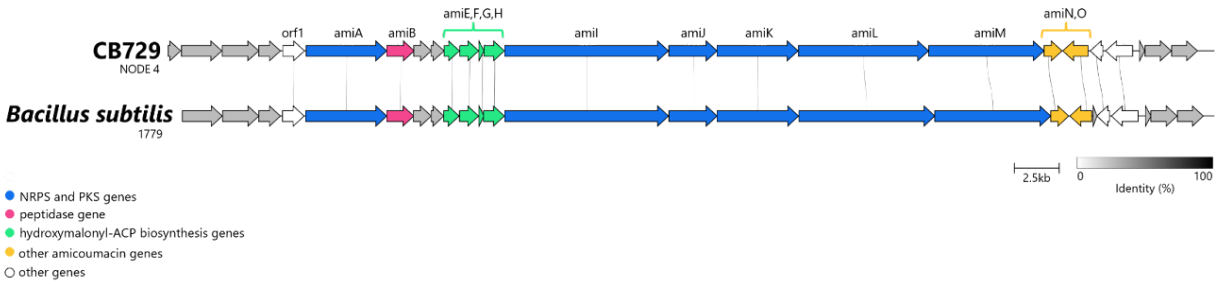


Figure 2. Clinker plot of CB729 Node 4 cov_60.218409 and *B.subtilis* 1779 with labels and colored legend based on the *B.subtilis* 1779 pathway.

Of the compounds in the amicoumacin family, *B.subtilis* 1779 primarily produces active amicoumacins, such as amicoumacin C.[6] In contrast, CB729 was observed to primarily produce the *N*-acetylated form of amicoumacin C. The structural differences between these two products were most likely caused by a *N*-acetyltransferase residing within the CB729 genome (Figure 3). This conclusion was founded on the idea that some amicoumacin producers, like *Xenorhabdus bovienii* Moldova, possess an *N*-acetyltransferase in their BGC as a part of an antibiotic resistance mechanism.[3] Like the name implies, the *N*-acetyltransferase will acetylate active amicoumacins, resulting in biologically inactive compounds. Looking at the cluster for CB729, in which the major component observed was an *N*-acetylated amicoumacin, there was no *N*-acetyltransferase identified by antiSMASH.

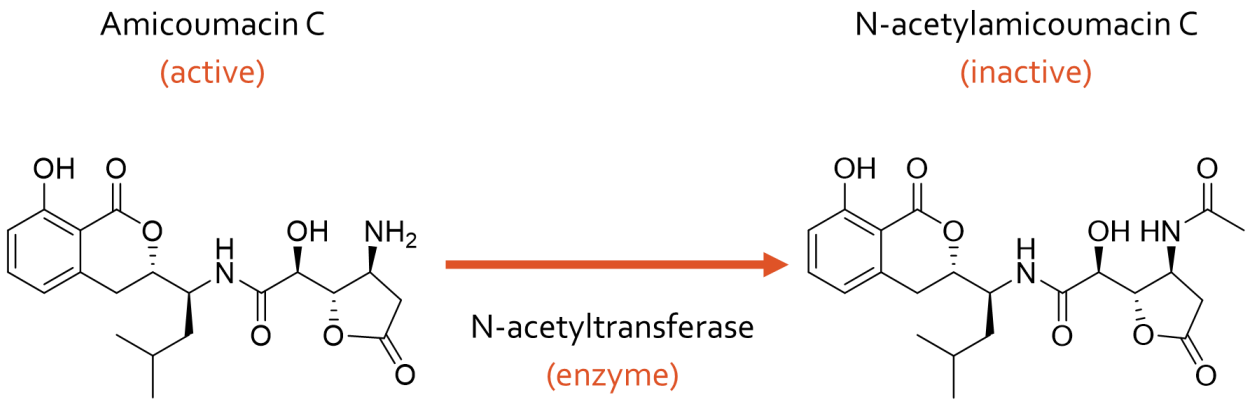


Figure 3. Illustration of the change in amicoumacin chemical structure most likely caused by a *N*-acetyltransferase enzyme.

Geneious tBLASTn Search

In the search for an *N*-acetyltransferase, Geneious software (http://www.geneious.com) was utilized to perform a tBLASTn of the entire CB729 genome against three acetyltransferase sequences from characterized natural product pathways. The three sequences used were an amicoumacin *N*-acetyltransferase from *X. bovenii*, an edeine transferase from *B. brevis*, and a paenilimycin transferase from *P. larvae*.[3,7,8] By running a BLAST search of CB729, 10 possible acetyltransferase sequences were identified and recorded, with the top hits having 30%, 24% and 26% ID for *X. bovenii*, *B. brevis*, and *P. larvae* respectively.

InterPro

The three sequences of known transferases from *X. bovenii*, *B. brevis*, and *P. larvae*, as well as the ten candidates identified from CB729 were analyzed for protein family membership using InterPro.[9] All sequences were recognized as part of the GNAT acetyltransferase YdfB-like family. Overall, analysis of family membership further supported the idea that the 10 sequences found though BLAST searches were good candidates for *N*-acetyltransferases within the CB729 genome.

IMG – Genome Database

The next step in this project was to identify a greater number of similar protein sequences in comparison to possible acetyltransferases in CB729. Identifying new sequences with a significant similarity was done through BLAST searches using the Integrated Microbial Genomes (IMG) system.[10] With this tool, gene hits associated with a search of the 10 sequences obtained from CB729 were exported as a FASTA file of results and a metadata file. This metadata file held information about each protein hit from IMG including genome name, product name, sequence, and other data provided by the database. Because IMG was used for obtaining sequences, the associated metadata was consistent, making later analysis of the SSN created easier.

EFI – Enzyme Similarity Tool

By compiling hits from IMG, those from BLAST, and literature searches, a total of 5,548 related gene sequences were uploaded to the EFI – Enzyme Similarity Tool.[11] The information generated was then analyzed through EFT-EST and used to generate a sequence similarity network (SSN). Specifications for sequence cutoffs and percent similarity values made through EFI for the creation of the SSN can be found in supplemental information (Figures 6,7)

Cytoscape Network

As a way to visualize similarities and differences between sequences based on their percent identities, a Cytoscape network was built from EFI data.[11] Figures 4 and 5 illustrate the full network with 100% and 55% identity for each node respectively and highlighted in each figure are the names of each literature sequence, as well as the name and genomic location of identified CB729 sequences included in the network.

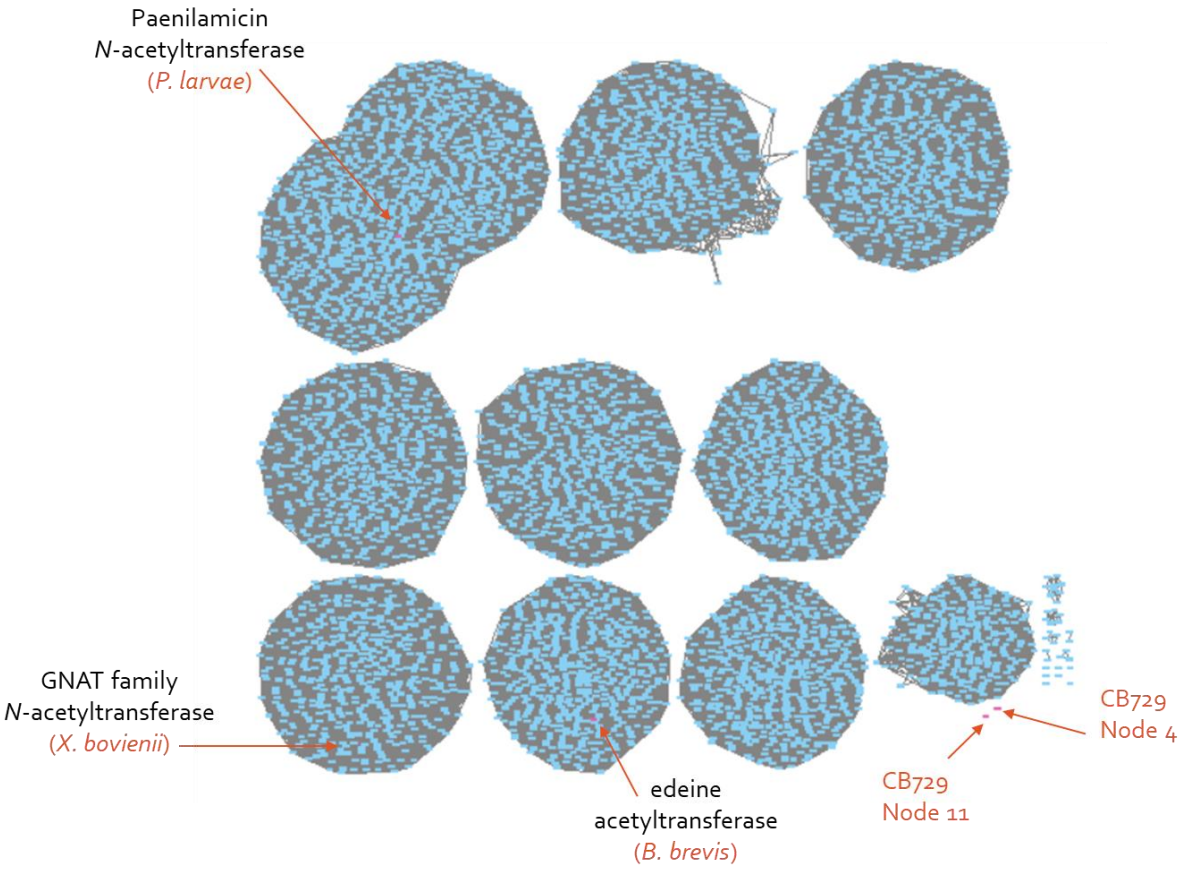


Figure 4. Full SSN with 100% identity of sequences with literature and CB729 sequences highlighted.

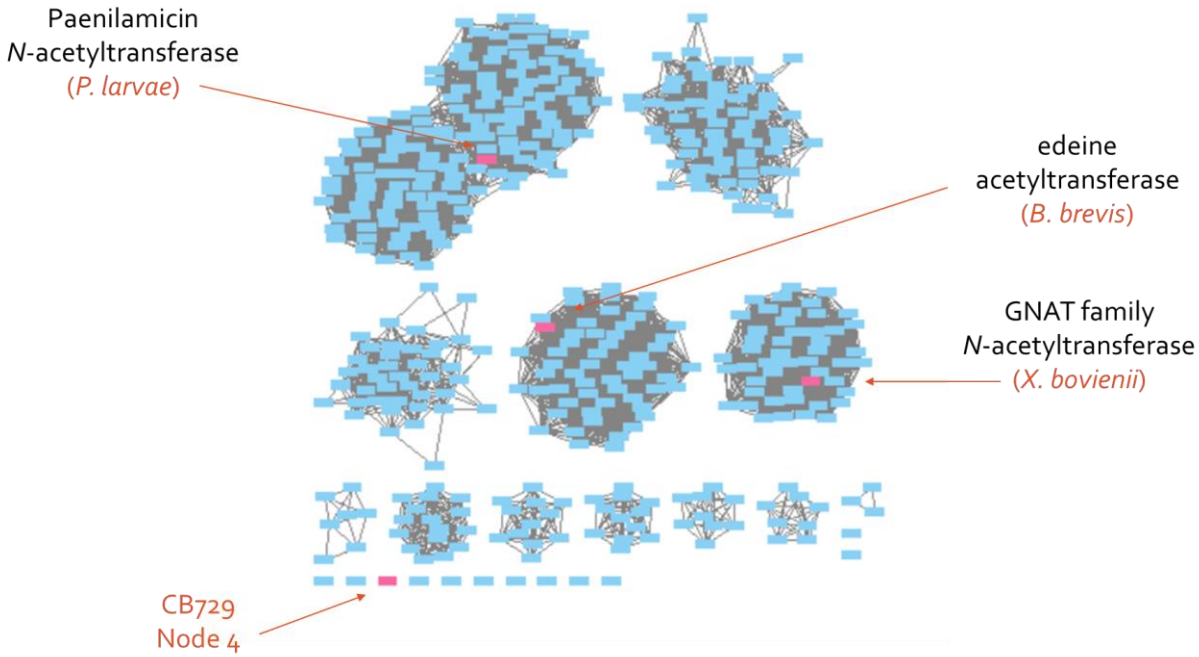


Figure 5. 55% identity SSN with literature and CB729 Node 4 sequences highlighted.

Results & Discussion

In the context of Figure 5, the possible acetyltransferase located in Node 4 of the CB729 genome remained isolated as its own node in the SSN. In referencing back to Figure 2, CB729 Node 4 was also the location of the BGC most likely associated with amicoumacin production. This identification supports the conclusion that there is a possible *N*-acetyltransferase residing nearby the piece of the genome which contains the producing BGC. Overall, this is a promising step towards identifying exactly which enzyme in CB729 is responsible for converting regular amicoumacins into *N*-acylated forms. Performing another BLAST search of the sequence of interest against the CB729 contig could determine the exact position of this sequence in the context of the genome. By utilizing multiple computational tools and the vast availability of genomic data, this project was able to make progress in understanding the production of a characterized natural product family by an uncharacterized citrus bacterium.

Future steps into this subject include investigating the genomic context of CB729 Node 4 and being able to specifically annotate the *N*-acetyltransferase gene most likely involved in the biosynthetic pathway. Similarly, by looking at the genomic context of the specific sequence, it could possibly determine what characteristics make this sequence significantly different from others in the same gene and protein family. Also, there is a large volume of information that can be inferred from looking at the various characteristics that group sequences together in the SSNs

produced. If there was a specific function or mechanism shared by sequences with a higher percent identity, there may be more conclusions that can be drawn about the characteristics of the identified sequence from CB729.

Supporting Information

For the creation of the SSN, minimum and maximum sequence length cutoffs were set between 140 and 378 and an alignment score threshold was set to 30%. These values were determined through graphical dataset analysis in EFI-EST (Figures 6,7). By finalizing the sequences through these thresholds, tiny sequences that were likely truncated or larger ones that might have corresponded to proteins of another family were removed from analysis. This resulted in a total of 5,258 sequences being used in the final network.

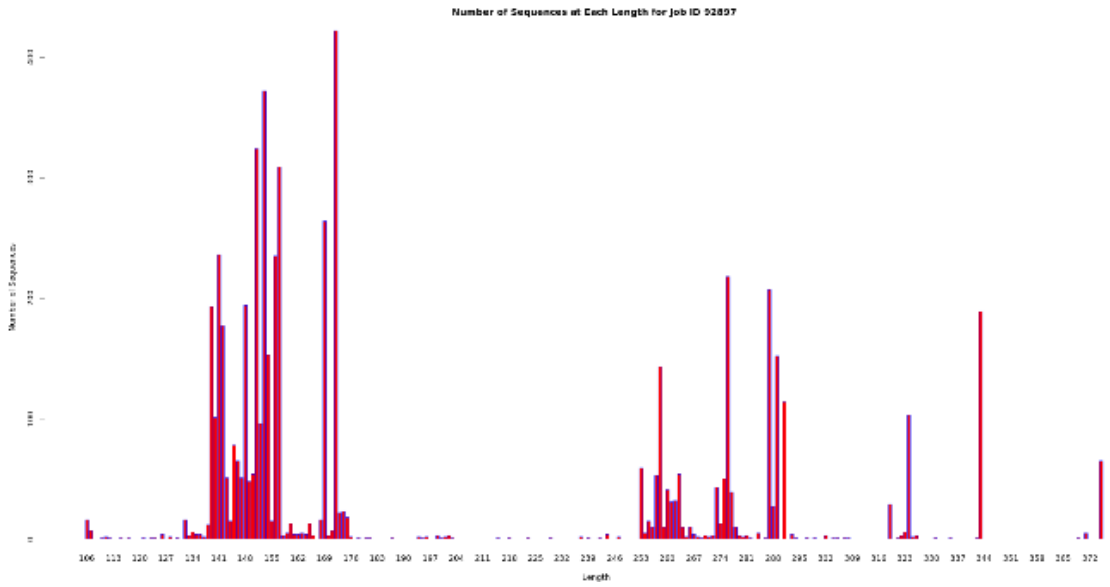


Figure 6. Histogram of sequences as a function of full length used in determining sequence length cutoffs for resulting SSN.[12]

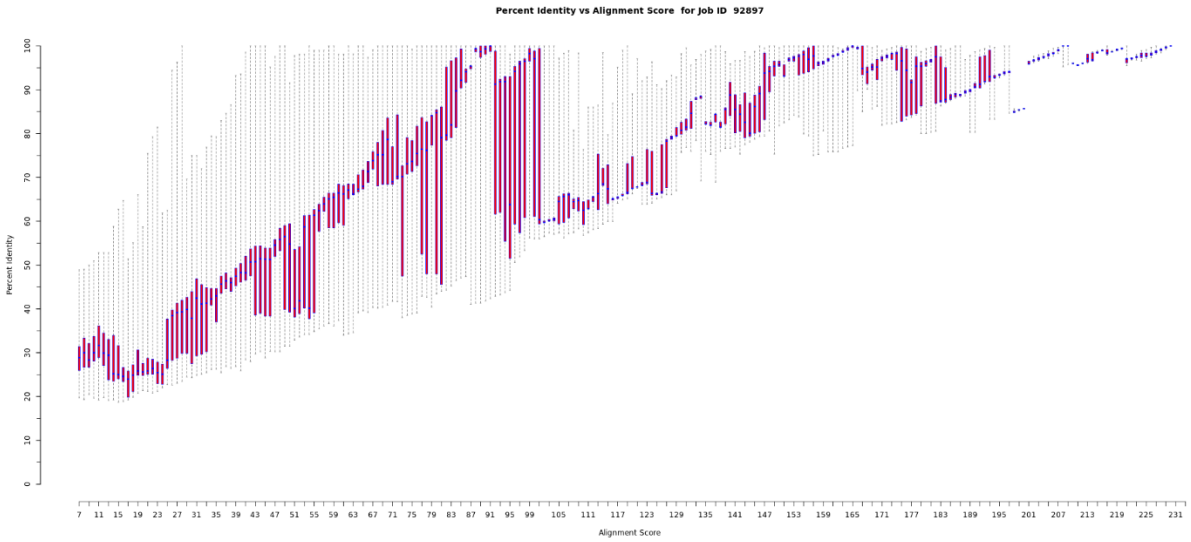


Figure 7. Percent ID vs. alignment score boxplot used in determining percent identity threshold for resulting SSN.[12]

Acknowledgements

I would like to extend my sincere thanks to the PLNU Chemistry Department for their guidance throughout this project, as well as my committee members, Dr. Ariane Jansma and Taylor Steele, for their generous support. I would also like to thank our collaborators at UC Riverside, Dr. Caroline Roper, Dr. Alex Blacutt, and Dr. Flavia Campos Vieira. Thank you to all of my fellow lab mates, Reese Bos, Corrie Fyle, Nathan Jenkins, Andrew Kamemoto, Olivia Owen, Heather Rainbow, Makenna Williams, and Lauren Wilson. Lastly, I would like to express my deepest gratitude to my mentor, Dr. Katherine Maloney, for her leadership, knowledge, and constant support.

References

1. Zhang, C. (2019). Citrus greening is killing the world's orange trees. Scientists are racing to help. Chemical and Engineering News. https://cen.acs.org/biological-chemistry/biochemistry/Citrus-greening-killing-worlds-orange/97/i23

2. Blacutt, A. et al. (2020) An In Vitro Pipeline for Screening and Selection of Citrus-Associated Microbiota with Potential Anti-"Candidatus Liberibacter asiaticus" Properties. *ASM Journals, 86(*8). https://doi.org/10.1128/AEM.02883-19

3. Park, H. B., Perez, C. E., Perry, E. K., & Crawford, J. M. (2016). Activating and Attenuating the Amicoumacin Antibiotics. Molecules (Basel, Switzerland), 21(7), 824. https://doi.org/10.3390/molecules21070824

4. Blin, K. et al. (2021). AntiSMASH 6.0: Improving cluster detection and comparison capabilities. *Nucleic Acids Research*, *49*(W1). https://doi.org/10.1093/nar/gkab335

5. Gilchrist, C. L., & Chooi, Y.-H. (2021). Clinker & Clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics*, *37*(16), 2473–2475. https://doi.org/10.1093/bioinformatics/btab007

6. Li, Y. et al. (2015). Directed natural product biosynthesis gene cluster capture and expression in the model bacterium Bacillus subtilis. Scientific reports, 5, 9383. https://doi.org/10.1038/srep09383

7. Westman, E. L., Yan, M., Waglechner, N., Koteva, K., & Wright, G. D. (2013). Self resistance to the atypical cationic antimicrobial peptide edeine of Brevibacillus brevis Vm4 by the N-acetyltransferase EdeQ. Chemistry & biology, 20(8), 983–990. https://doi.org/10.1016/j.chembiol.2013.06.010

8. Dang, T. et al. (2022). Molecular basis of antibiotic self-resistance in a bee larvae pathogen. Nature communications, 13(1), 2349. https://doi.org/10.1038/s41467-022-29829-w

9. Paysan-Lafosse, T. et al. (2022). Interpro in 2022. *Nucleic Acids Research*, *51*(D1). https://doi.org/10.1093/nar/gkac993

10. Chen, I.-M. A. et al. (2022). The IMG/M data management and analysis system V.7: Content updates and new features. *Nucleic Acids Research*, *51*(D1). https://doi.org/10.1093/nar/gkac976

11. Shannon, P. et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303

12. Zallot, R., Oberg, N., & Gerlt, J. A. (2019) The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry 2019 58 (41), 4169-4182. https://doi.org/10.1021/acs.biochem.9b00735