Using Gene Neighborhoods to Mine for Bacteriocins

Md Nafiz Hamid[1,2], James T Morton[4], Stefan D Freed[3], Shaun W Lee[3], Iddo Friedberg[2]

[1]Program in Bioinformatics and Computational Biology, [2]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University

[3]Department of Biological Sciences, University of Notre Dame, [4]Department of Computer Science and Engineering, UC San Diego

What are Bacteriocins?

Bacteriocins are peptide-derived molecules produced by bacteria that function as virulence factors, signaling molecules, and antimicrobials. They are surrounded by context genes that are responsible for the translation, modification, transport and self-immunity from the bacteriocin.

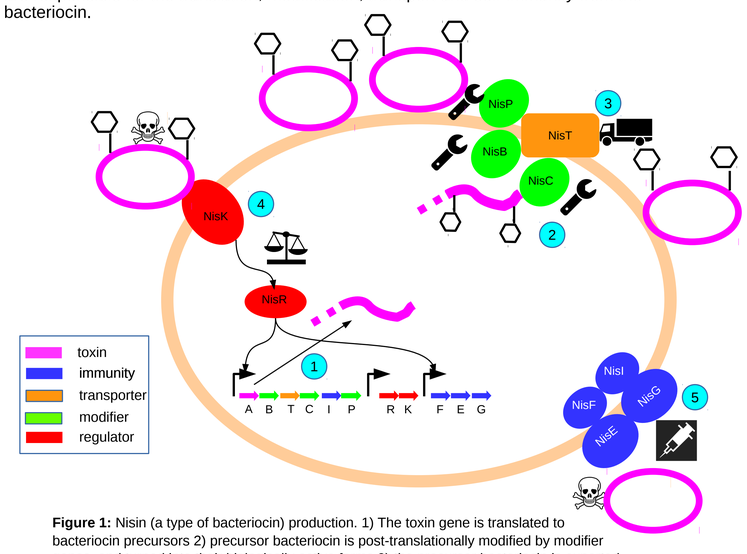


Figure 1: Nisin (a type of bacteriocin) production. 1) The toxin gene is translated to bacteriocin precursors 2) precursor bacteriocin is post-translationally modified by modifier genes, and turned into their biologically active forms 3) the precursor bacteriocin is exported by transporter genes 4) regulator genes control the production of bacteriocins 5) immunity genes protect the bacteria producing the bacteriocin from the toxin. These context genes have been shown to be largely conserved across unrelated species. As bacteriocins may be non-homologous, finding context genes might lead to finding bacteriocins.

The Bacteriocin Operon and Gene Block Associator (BOA) Pipeline:

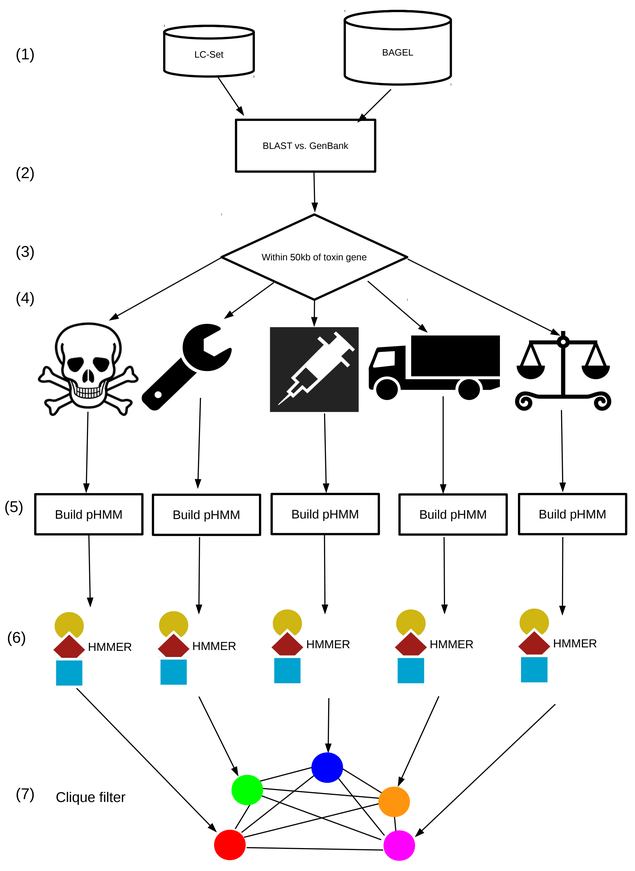
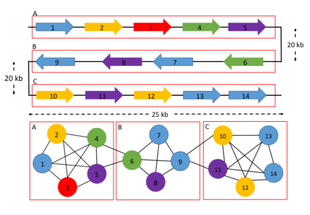


Figure 2: An overview of the BOA Pipeline. (1) Construction of the LC-Set(**L**iterature **C**urated data set – seven experimentally identified bacteriocin associated gene blocks) and the BAGEL(another bacteriocin database) set; (2) BLAST LC and BAGEL genes against all bacterial & archaeal genomes e-value = 10−5; (3) select the ORFs within ±50 kb of homologs to toxin genes (4) assign ORFs to one of the following classes (left to right): toxin, modifier, immunity, transport, regulation; (5) build pHMMs from each category: cluster sequences using CD-HIT, align sequences in each cluster using MAFFT, then use hmmbuild from the HMMER suite to construct HMMs; (6) run hmmsearch from the HMMER suite against the genome files to extract more sequences from each category, remove predicted false positives using a threshold score as explained in Methods (7) use a clique filter to identify genes that are close together



Clique Filter:

- Each node is a gene
- If two genes are within 25kb of each other, create an edge
- Clique containing all of the functions are predicted to be operons.

Results

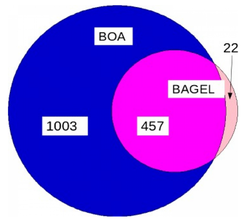


Figure 3: 95 % (457 out of 479) of BAGEL toxins were predicted by BOA. BOA predicted an additional 1003 toxins throughout bacterial genomes that are not listed in BAGEL. Twenty-two BAGEL toxins were not predicted by BOA

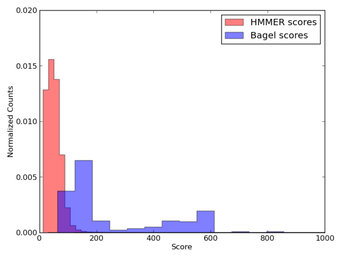


Figure 4: Determining the threshold for similarity-based search of toxin genes. To determine an adequate threshold for inferring homology, we examined the distribution of HMMER scores for homologs for predicted toxin genes (red) and BAGEL-derived toxin genes (blue). BAGEL toxin gene scores were used to set a minimum threshold of acceptance for HMMER scores for predicted genes.

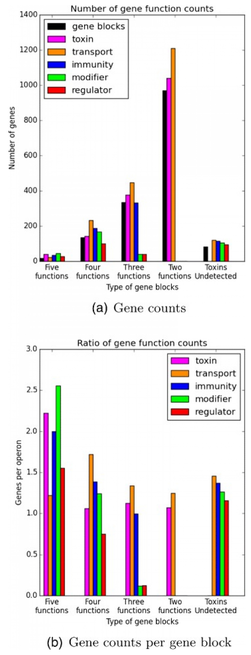
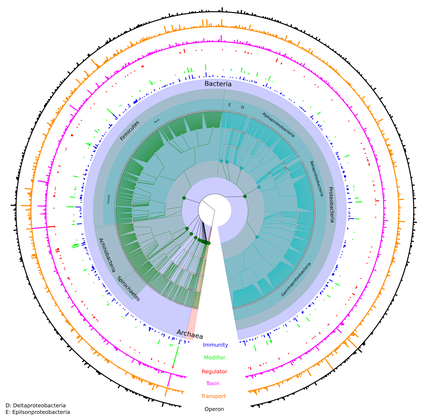


Figure 5: Gene blocks were classified by the number of detected functions. **(a)** number of total genes found **(b)** gene counts per detected block

Bacteriocin species tree



Mining for new context genes using keywords

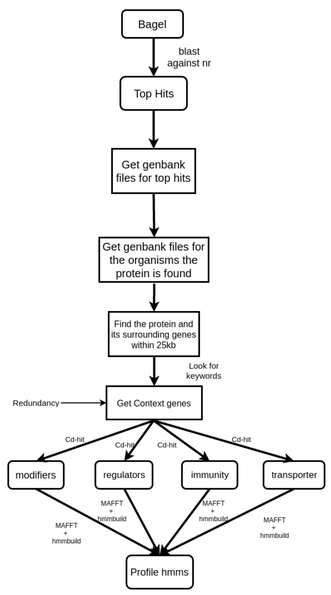


Figure 6: Pipeline for finding new context genes. We find the locations of the bacteriocins from BAGEL within their organisms, and try to find surrounding putative context genes through keywords (i.e. 'modifier', 'transport', 'regulator', 'immunity' and others) in their annotations. We find the unique genes among them through CD-HIT, and then build profile HMMs from them.

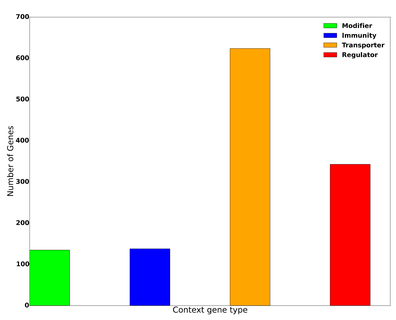


Figure 7: Count of newly found context genes to build new pHMMs.

Mining OperonDB to find new bacteriocins

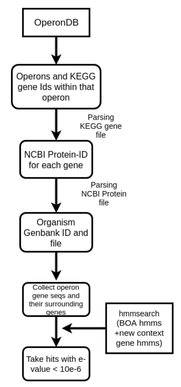


Figure 8: Pipeline for applying our pHMMs to bacteria operons in operondb. We collect all operons, and their surrounding genes, and then apply BOA pHMMs as well as the new pHMMs to these genes to find new bacteriocins and context genes.

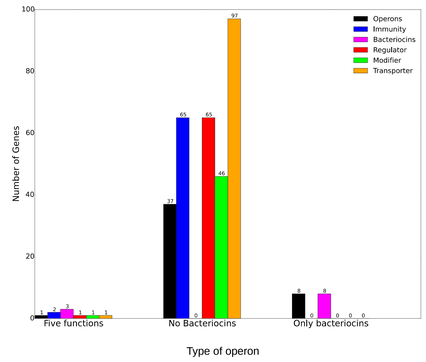


Figure 9: Results from applying our profile hmms to operondb. Between overlapping operons, we only took operons with the highest number of genes associated in operondb. The most interesting cases, operons with bacteriocins and all context genes, operons with all 4 context genes and no bacteriocins, and operons with only bacteriocins are shown above.

Conclusions

- **BOA** is the first time a curated data set has been established for bacteriocin context genes.
- **BOA** was able to identify the majority of bacteriocin gene clusters that **BAGEL** identified.
- **BOA** also predicted over seven times more bacteriocins in whole bacterial genomes than **BAGEL**, including many identifiable bacteriocin gene blocks with experimental validation.
- From operondb, after applying our profile HMMs, we could identify 37 operons, that have all 4 context genes, but no bacteriocins. These operons are likely candidates for new undetected bacteriocins.
- The **plnJKLR** operon in operondb, identified by us to have (together with the surrounding genes) all four context genes and bacteriocins, is found experimentally verified to be an operon with bacteriocin producing genes.
- We have identified 8 operons where only bacteriocins were found, but no context genes were found. These might have context genes we could not detect.

Future Work

- **BOA** encompasses a large number of taxa, the information in **BOA** can be used to explore the evolutionary development of bacteriocin gene blocks and how different biosynthetic loci have evolved in different clades.
- Find novel computational ways to detect bacteriocins and bacteriocin operons more efficiently.

References

- James T Morton, Stefan D Freed, Shaun W Lee and Iddo Friedberg. **A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins**. BMC Bioinformatics 201516:381
- Anne de Jong, Auke J. van Heel, Jan Kok and Oscar P. Kuiperds. **BAGEL2: mining for bacteriocins in genomic data.** Nucleic Acids Research (2010).

More

Paper

Code

Contact

{nafizh, idoerg}@iastate.edu

This study was funded, in part by NSF award ABI Innovation 1551363 awarded to IF