Building a Peptide Database to Perform Automated Forensic Analysis of Toxins using Mass Spectrometry

Capstone Report, Masters of Science, Bioinformatics

Advisors: Dr. Eckart Bindewald, Dr. Daniel Sierra-Sosa, Dr. Miranda Darby

Daniel Vogel, Hood College

2022-12-10 12:57:58

Abstract

Bioterrorism has created a need for the rapid analysis of samples which may contain toxins, viruses, or other deadly agents. Mass Spectrometry (MS) provides a tool for use in proteomics for accurate and comprehensive profiling of proteins. Automated searching of MS sample data for matches against a proteome database is useful for forensic analysis of samples and an effective countermeasure to Bioterrorist attacks. A software tool called MARLOWE was tested and it worked well for MS sample analysis but failed to identify organisms of interest such as the toxin, Arbin, that were missing from the KEGG.JP (Kanehisa et al., 2002) database on which it relies. FTP access to KEGG.JP is cost prohibitive and the database creation process is compute and time intensive. Here we create a database for MARLOWE using publicly available proteomes from the Uniprot.org database (Consortium, 2020). By creating a process to update this MARLOWE database, we can ensure that target organism are present, even as newer organisms are added to UniProt. I have also created a new code base which works on Linux so that this pipeline can run on High Performance Computing clusters.

Contents

Introduction

Bioterrorism

Bioterrorism is the use of microorganisms or toxins by terrorist or extremists' groups to produce weapons which cause death and disease (Jansen, Breeveld, Stijnis, & Grobusch, 2014). The use of biological agents (Bioweapons) to cause harm or death is not a new concept; countries have been engaging in bioterrorism for hundreds of years. Infectious diseases were recognized for their potential impact on people and armies as early as 600 BC. The crude use of filth and cadavers, animal carcasses, and contagion had devastating effects and weakened the enemy. Polluting wells and other sources of water of the opposing army was a common strategy that continued to be used through the many European wars, during the American Civil War, and even into the 20th century (Riedel, 2004). More recently, an attack using anthrax-laden letters mailed to media organizations and politicians (Pal et al., 2017). Progress made in biotechnology and biochemistry has simplified the development and production of such biological weapons (Riedel, 2004), thus it is predicted that proliferation of Bioweapons will increase in the next decades. Bacterial and plant protein toxins are among the most powerful poisons known and are considered as potential agents used for bioterrorism and warfare (Duracova, Klimentova, Fucikova, & Dresler, 2018). Biodefense strategies which include early and accurate threat detection are essential in mounting a successful response to bioterrorism.

Mass Spectrometry

Mass spectrometry (MS) can be utilized in all stages of such a response: from Bioweapon detectors to accurate forensics classification for successful prosecution (Demirev & Fenselau, 2008). MS is currently the most comprehensive method for the quantification of proteins (Sinha & Mann, 2020). It provides a valuable tool for forensic sample analysis. The accuracy of MS can also distinguish between similar compounds and organisms such as the castor bean vs the derivative toxin, ricin. Current methods, involving manual database searches, are slow, and depend on expert knowledge. Automated solutions produce effective results for non-experts and narrow the search field for experts to perform more in-depth analysis. We must refine software and methods, improving accuracy and efficiency to combat the Bioweapon threat.

Bottom-up Proteomics

Here we focus on bottom-up proteomics where proteins are digested into smaller peptides, which are analyzed by MS. MS detect the presence and quantity of peptides using properties of mass and net-charge. Mass spectrometers can only analyze gaseous ions, therefore peptides are converted into peptide ions which are separated by their mass-to-charge ratio (m/z).

Samples are prepared for MS by digesting the long polypeptide chains using protease enzymes such as Trypsin, Chymotrypsin or Pepsin. These break proteins into smaller peptides which can be measured by MS. Protease digestion acts to normalize and compartmentalize the biochemical heterogeneity of proteins within a sample as peptides and may create a less heterogeneous mixture when protein splice isoforms and post-translational modifications are considered (Zhang, Fonslow, Shan, Baek, & Yates III, 2013). One consideration due to the sensitivity of the MS analysis is that contaminant proteins will appear from Trypsin and human interaction and must be removed from analysis. Contaminants include human skin or animal proteins from the digesting enzyme.

Bioinformatics Analysis

Bioinformatics analysis of MS output involves comparing matching the spectral output data from a sample with the theoretical spectra predicted from known genome or amino acid sequences contained in databases (Pere, 2020). Amino Acid (AA) sequences of the peptides are determined using PEAKS® or Novor de-novo sequencing software. The MARLOWE software tool, developed at the Pacific Northwest National Lab (PNNL), to automate this database searching, producing a list of potential candidate compounds or organisms and likelihood scores for each candidate. MARLOWE is an R-Markdown application where the user adjusts a few input parameters to control sensitivity and set sample data input file names. It then performs analysis and generates documents to show the candidate scoring. In testing the MARLOWE search with a sample of the toxin, abrin, no matches were found as this compound was missing from the database.

We will rebuild the MARLOWE analysis software database from UniProt, a publicly available data source. The new database will be generated via in-silico digestion of Amino Acid sequences found in UniProt, using Trypsin. This will result in more accurate and updated searches. The MARLOWE database can be updated, at no cost as more samples are added to UniProt. We will evaluate the success and propose a method to process lab samples digested with enzymes Chymotrypsin or Pepsin (Dau, Bartolomucci, & Rappsilber, 2020).

Literature Review and Previous work

MARLOWE software tool

MARLOWE compares these peptides to a MySQL database which has been constructed from the KEGG.JP protein database (Kanehisa et al., 2002). The construction of the database requires taking the long amino acid sequences from each organism and digesting in-silico using an R subroutine, into smaller peptide sequences that match the digested lab sample. The database contains organism, protein, and peptide tables. MARLOWE then uses an algorithm which compares the peptides from MS/PEAKS® data with the database peptides and makes a list of candidate organisms that may be contained in the sample. It produces scores based on the quantity of matches and a heat map showing the most likely organisms contained in the sample data.

MARLOWE algorithm

The peptides from a sample are searched against those from other organisms and strong-peptides are tagged. These have more value to the search algorithm since multiple hits is a better indicator that the match was not random. Each organism with peptide matches is scored and a list of potential candidate organisms is produced along with a table of peptides that matched.

Specific Aims

Aim 1 Create R package to parse UniProt FASTA

Aim 1 is to create a new R package which will extract required fields from publicly available UniProt data (Consortium, 2020) in fasta format. A major weakness of the MARLOWE tool is that the database was built using protein data from KEGG.JP, more than 2 years ago. The database cannot be updated without a paid subscription to the KEGG.JP FTP site which costs $25,000 per year for a single user commercial license. The KEGG data extraction subroutines in MARLOWE rely on regular expression to parse human readable .ent files from KEGG.JP. The .ent files are proprietary to KEGG so these subroutines are not useful for any other database sources (Csordas et al., 2012). The fasta format is a standard so the R data extraction package will be adaptable to UniProt or other protein databases where fasta output is available. Successful adoption of the UniProt data will make MARLOWE cheaper and more accessible for use as licensed access to the KEGG.JP FTP site is not required.

I created a new UniProt based peptide database by downloading data from Uniprot's reference proteomes in fasta format and parsed it via the parse_fasta package to extract the data fields that MARLOWE requires. Creating the MARLOWE database is a computationally intense process taking weeks on a powerful workstation or server. There are over 22,000 organisms. By extracting this data into a separate file for each, I can develop a faster method to create the database using parallel processing on a High Performance Computing (HPC) cluster. I have ported the MARLOWE code and database from Windows to Linux to make this possible.

Aim 2 Create a MARLOWE database with UniProt Proteomes

Aim 2 is to improve the accuracy and performance of the MARLOWE tool by updating the database with new samples and creating an efficient process to update the database as needed. When NBACC scientists tried to perform an analysis of a sample containing the toxin "abrin", the program did not identify abrin because the database was out of date and did not contain entries for abrin. These were added to the KEGG.JP database in the year after MARLOWE was created. Abrin protein sequences are currently contained in both UniProt and KEGG.JP protein databases. These databases are updated constantly with protein sequences, often in response to a new toxin or virus of interest. Addition of the latest entries from UniProt will improve the accuracy of the MARLOWE pipeline, increasing the chance to identify the candidate toxin without manual searching.

Aim 2. I will download subsets from UniProt by using queries which will contain the compounds in the test samples, milk, abrin, and castor bean. I will then compare the resulting MARLOWE/UniProt analysis to that done with the MARLOWE/KEGG to validate the sample was found and compare the scoring. This will lead to a process which can be used to update the database as needed, in the future. As more entries are added to UniProt

or other protein databases are identified, this process can be repeated. After validation of the extraction and database creation, we can develop a repeatable procedure to download the entire UniProt database. Additionally, if I can obtain access to the current KEGG.JP FTP site, I will create an updated MARLOWE/KEGG database to validate that “abrin” can now be located using the original MARLOWE code.

Resources, Tools, and Research Methodology

Hardware and Software Versions

I installed MARLOWE on a Linux server with a 32-core AMD CPU and 500G RAM. The storage for the database was a high speed NVME SSD. The Operating System was Ubuntu 20.04LTS. I have also installed it on a MS Windows 10 workstation with 18-core Intel CPU and 256G RAM. MARLOWE was designed to run with Microsoft Windows® but I have converted it to run on Linux so that I may use HPC resources. I will test it on Ubuntu 20.04LTS and CentOS 7. The conversion to Linux involved replacing some of the R packages used for File selection with code that will work on Linux. The KEGG database mysqldump (backup file) had to be edited to change some of the collation utf8mb4_0900_ai_ci used by MySQL 8.0 which is not supported by MariaDB. It was replaced by collation utf8mb4_unicode_520_ci. This collation works on MySQL 8.0.31-0ubuntu0.20.04.1.

MARLOWE source packages were forked into a GitHub.com repo so I can create new packages which can be run on Linux platforms. The code is written with R version 4.2.1 using the RStudio IDE and a database from UniProt data will be created with MySQL version 8.0.31 for Ubuntu Linux and MariaDB 5.5 for CentOS 7 Linux. Collation utf8mb4_unicode_520_ci is not supported on MariaDB 5.5 so if an existing MS Windows MARLOWE KEGG database is restored on Linux, it must be changed. This issue presented itself as I ran the data samples on the Linux platform using both KEGG and UniProt databases for comparison.

I validated the output of MARLOWE (current version MySQL 8.0 on Windows 10, R version 3.6.1) with 8 data files from biological samples including Fish, Milk, Oyster, Juice and Castor bean. To perform the analysis of the two databases, MARLOWE was run against each database with the same samples. The output from each was compared with the actual contents of the sample in order to make a conclusion on performance or improvements needed.

R packages Developed for Parsing FASTA files

The FASTA format for the UniProt database contains the minimum required fields for MARLOWE but must be parsed differently since it is vastly different from the KEGG .ent format which is more suited for human readers. There are in fact different FASTA header formats used by UniProt Uniref and UniProtKB. The UniProt Proteome FASTA files use the UniProtKB format. I wrote a parse_fasta R function, which examines the file to determine which header format is in use and applies the appropriate parsing via Regular expressions. Here is an example of the fields extracted from the proteome FASTA headers that were used for the organisms inserted into the database. UniProt identifies organisms with NCBI taxonID (OX). The KEGG database used the unique kegg_id identifier. For our purpose, I am prefixing “U” in front to the integer taxonID to replace the kegg_id for organisms downloaded from UniProt. In the example below, R. communis taxonID = 3988 so the kegg_id =

U3988 when inserted into the candidate database. The KEGG .ent files provided more fields than are available in the header so many fields are set to NULL for the UniProt derived organism data. These include Pathways, Database Links, Module, Brite, Position, and Motif. They are not required for the MARLOWE algorithm.

UniProt Release 2022_04

```
UniProtKB Fasta header format

  These files, composed of canonical and additional sequences, are non-redundant
  FASTA sets for the sequences of each reference proteome.

  For further references about the standard UniProtKB format, please see:
   http://www.uniprot.org/help/fasta-headers
   http://www.uniprot.org/faq/38

 >db|UniqueIdentifier|EntryName ProteinName OS=OrganismName OX=OrganismIdentifier [GN=GeneName ]PE=ProteinExistence SV=SequenceVersion

 Example from the R. communis Proteome FASTA file of headers used for 5 protein sequences

 >tr|B9R7K7|B9R7K7_RICCO Cytochrome P450 OS=Ricinus communis OX=3988 GN=RCOM_1592680 PE=3 SV=1
 >tr|B9R8T7|B9R8T7_RICCO Cinnamoyl-CoA reductase, putative OS=Ricinus communis OX=3988 GN=RCOM_1602080 PE=4 SV=1
 >tr|B9R9L1|B9R9L1_RICCO 3-ketoacyl-CoA synthase OS=Ricinus communis OX=3988 GN=RCOM_1498550 PE=3 SV=1
 >tr|B9RAM4|B9RAM4_RICCO 26S proteasome non-atpase regulatory subunit, putative OS=Ricinus communis OX=3988 GN=RCOM_1507340 PE=4 SV=1
 >tr|B9RBT9|B9RBT9_RICCO C-4 methyl sterol oxidase, putative OS=Ricinus communis OX=3988 GN=RCOM_1681040 PE=3 SV=1
```

Results

UniProt candidate database

I built a minimal "candidate" database with UniProt (Release 2022_04) Proteome data for 9 organisms including those contained in the lab samples that were analysed with Mass Spectrometry. Building the database involves downloading and parsing fasta proteome files, then inserting into the database along with the amino acid sequences for all proteins and the peptides that result from digesting these proteins with Trypsin. This process is shown in Figure 1. The parse_fasta function performs this logic.

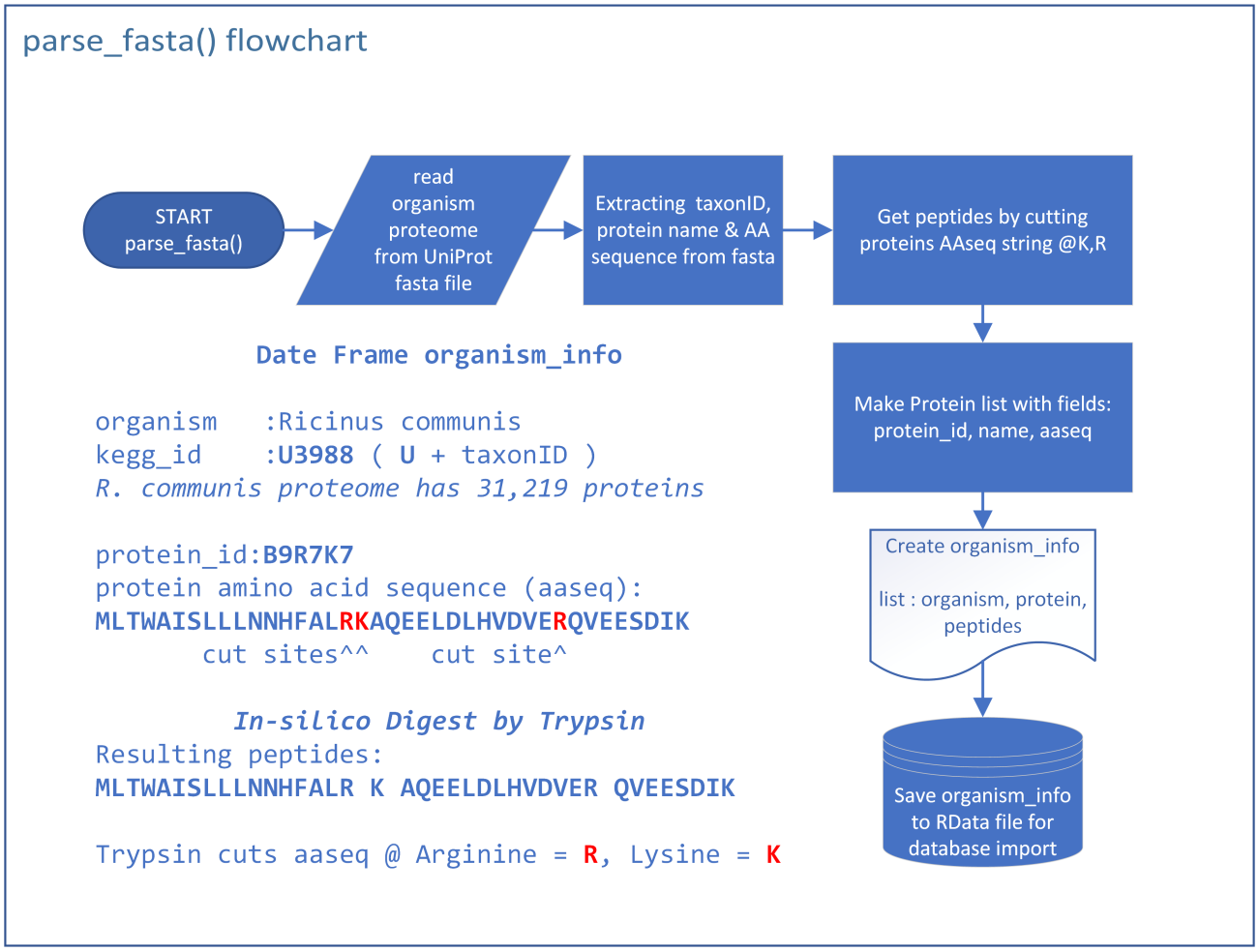

Figure 1: Flowchart showing parse_fasta in the database creating process. This function reads fasta files and produces .RData files as output which contain the elements of organism_info. These .RData files are used in buidling the database via the init_database function.

A final step is to upload NCBI taxonomy data for all organisms used to produce the MARLOWE heat maps. The following shows the contents of the MySQL database to verify organisms have been inserted correctly. As a validation of the success, the following SQL query shows the organisms inserted and they quantities of peptides

which resulted from the in-silico digestion. The filtering for strong peptide count can also be verified. Strong peptides are defined as those which can be found in a genus. There is another definition of strong peptides used in sample analysis. In the analysis stage, strong peptides are multiple peptides found in the sample which match a single protein. The peptides are more reliable indicators of the presence of the protein than cases where a single peptide matches (Zhao & Lin, 2010).

Table 1: Query of Organisms Inserted into the Candidate Database

X	name	kegg_id	taxon_id	protein_count	peptide_count	strong_peptide_count
1	Bos taurus	U9913	9913	23844	652649	554285
2	Citrus clementina	U85681	85681	24934	586056	482274
3	Citrus sinensis	U2711	2711	28128	572368	471745
4	Crassostrea gigas	U29159	29159	25998	687216	624744
5	Crassostrea virginica	U6565	6565	33719	876976	805672
6	Ricinus communis	U3988	3988	31219	630447	520219
7	Pseudomonas fragi	U296	296	4324	85668	76797
8	Salvelinus namaycush	U8040	8040	35973	696618	599782
9	Chlamydia pneumoniae	U83558	83558	1052	23031	20258

The MARLOWE software was run with PEAKS output from 8 datasets which contained organism that are present in the UniProt database. Ricinus communis was contained in 4 of the samples and the other 4 were random samples created for testing to include milk, orange juice, oyster, and fish. MARLOWE was able to detect the sample with a strong signal score for 7 of the tests.

Heat maps produced as the final output by MARLOWE for each sample clearly show the presence of the target organisms contained in each sample. Scores were high when a target organism was identified by the algorithm. These heat maps are shown in the Figures chapter.

In the Orange Juice sample, the 2 Citrus organisms scored high, (130, 62) which is to be expected. There was also a presence of R. communis (28) detected which would be a false positive likely caused by overlap of peptides between these organisms. Further investigation is needed to determine the cause.

Poor results where evident in 555558-DeNovo, a R. communis sample, where B. taurus (4) and S. namaycush (3) both scored slightly higher than R. communis (2). However these low scores are all below the threshold to be considered as a match. Comparison with KEGG output for Sample 555558-DeNovo indicates that there is an issue with this sample's data. The KEGG output did not identify R. communis as a candidate for sample 555558-DeNovo.

Table 2: Comparing Results from UniProt and KEGG

Sample	UniProt	KEGG	Next_UniProt
R. communis 9	164	111	1
R. communis GC4	654	476	1
55551-DeNovo	420	303	1
555558-DeNovo	2	0	3
Fish-DeNove	330	16	302*
Juice-DeNovo	130	423	62**
Milk-DeNovo	432	109	4
Oyster-DeNovo	714	306	10

- Score 302 was for P. fragi bacteria which indicates spoilage. ** Score 62 was for another orange detected since the orange juice was a blend of 2 oranges. "'

The performance with UniProt data shows that the database creation is working for this small set of organisms. The scores for the UniProt hits are on average higher than those from KEGG, showing confidence. The KEGG database has more than 5,000 organisms. There are approximately 22,000 reference Proteomes on UniProt. We will create a more complete database with 5,000 to 20,000 UniProt Proteome for a fully functioning MARLOWE. This should lead to hits for the larger set of organisms.

Discussion

The process to build the database appears to be working but it is very slow on my test with a single server. There are 2 main steps: 1. Convert the UniProt proteome fasta files into RData and 2. Import a directory containing the Rdata files into MySQL.

We can convert the 22,000 UniProt Reference Proteomes into RData files. We could convert a subset of Proteomes. The proteome fasta are organized in the groups [Archaea, Bacteria, Eukaryota, Viruses] so that we could convert subsets of Virus, Bacteria, and Eukaryota but not Archaea. We could also select organisms based on taxonomy or other attributes such as presence of toxins. If we have a list of organisms to be inserted into the database, we can convert these first.

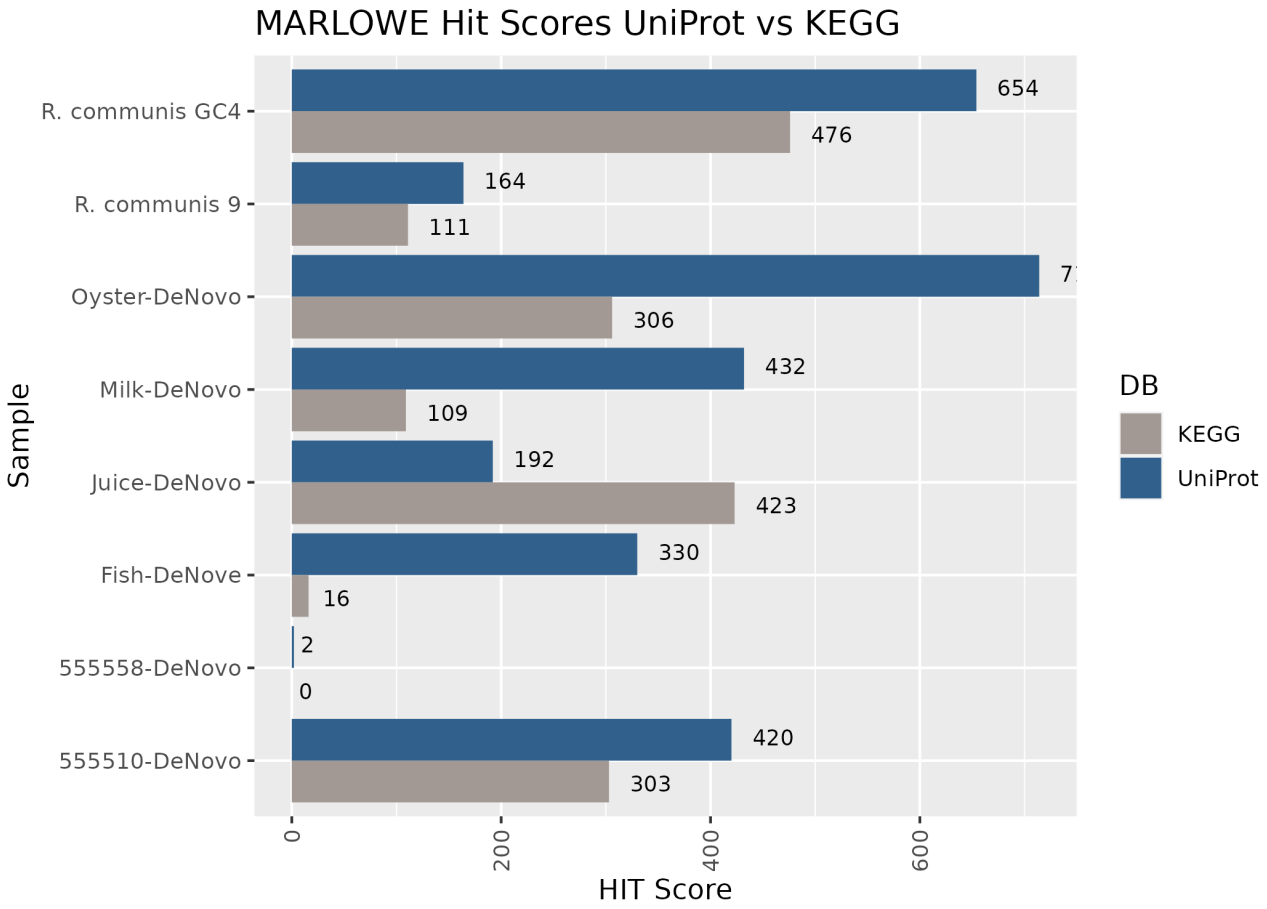

Figure 2: Visualization of results. It can be clearly seen that the UniProt database search resulted in highter scores than searches agains the KEGG database. The scores are the number of peptides in the samples that were found in the proteome for the organism after importing into the database. Overall UniProt correctly indentified the present of the sample in 8/8 case and KEGG in 7/8 of the cases tested.

Summary and Future Work

Alternate digestion enzyme

Currently MARLOWE only supports Trypsin digest. We can construct another version of the database where the peptides have been digested with an alternate protease enzyme (Dau et al., 2020).

Efficiency Improvements

The time required to build the sample database was about 24 hours with 9 organisms. We will need to improve the speed of this process using parallel computing and multiple servers in order to create the database. Exploring faster algorithms may also lead to improvements. I will also attempt to insert organisms from UniProt proteomes into the existing KEGG database. This would be an efficient way to append to data specific organism which match samples being tested.

User Interface Improvements

GUI interface would be possible by creating an R-Shiny version where the scientist could select their input files using a GUI and then the pipeline would run automatically and produce and output report that could be viewed and downloaded. Automated database updates via API would allow adding new organisms to the database by pulling from UniProt via API and inserting into the MARLOWE candidate databases.

Figures

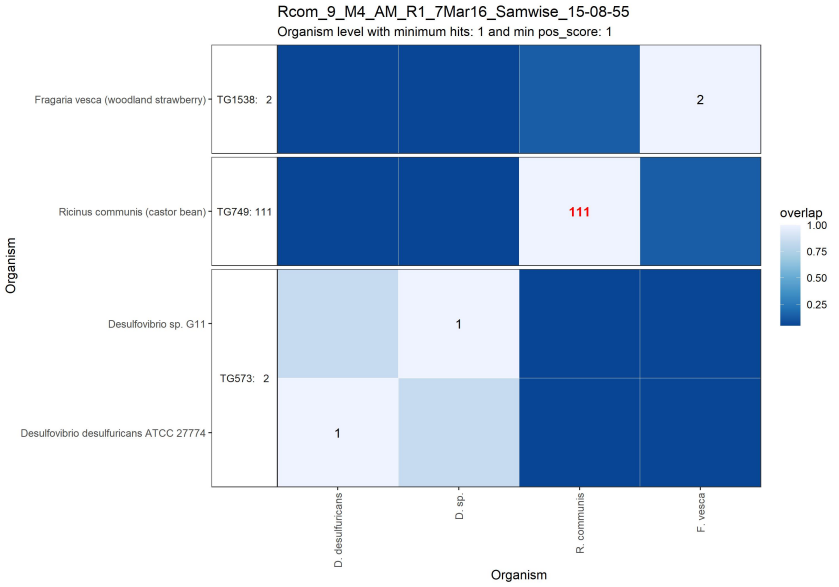

Figure 3: KEGG Heat Map R. communis (castor bean),Rcom_9 Sample

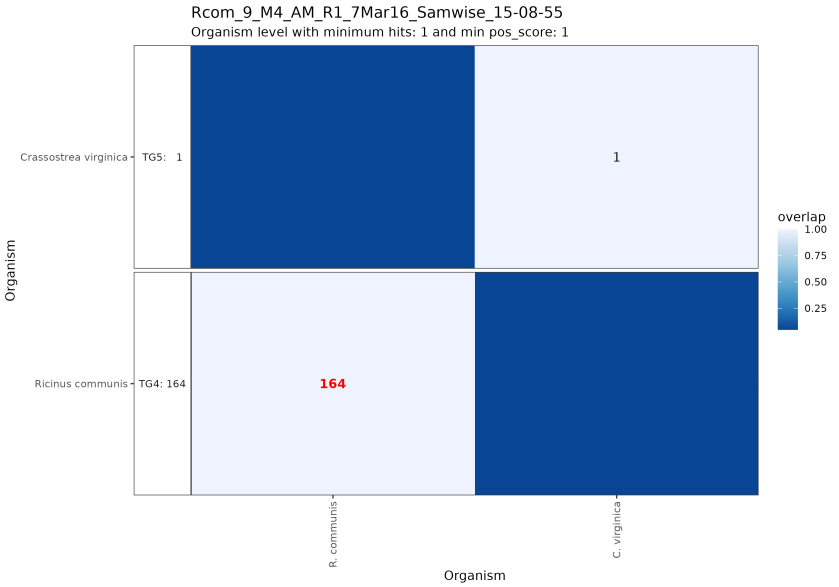

Figure 4: UniProt Heat Map R. communis (castor bean),Rcom_9 Sample

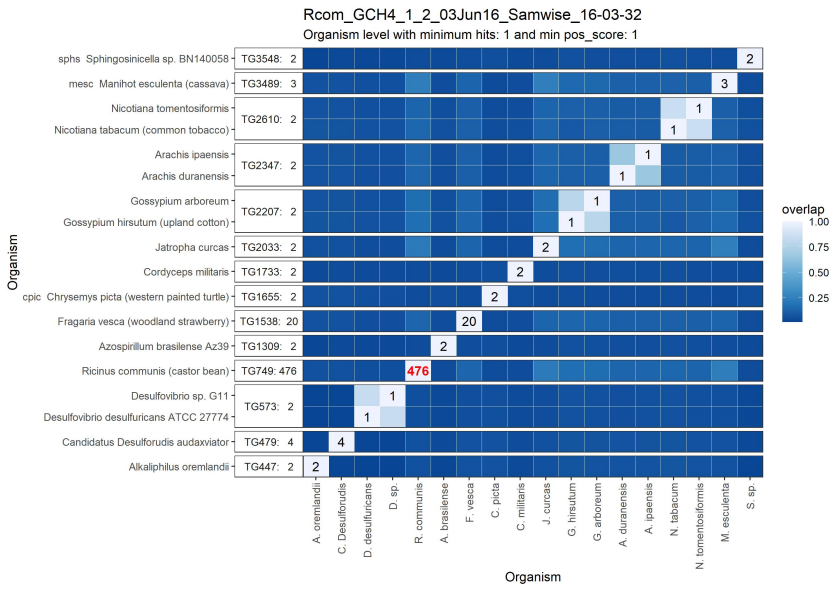

Figure 5: KEGG Heat Map R. communis,Rcom_GC4 Sample

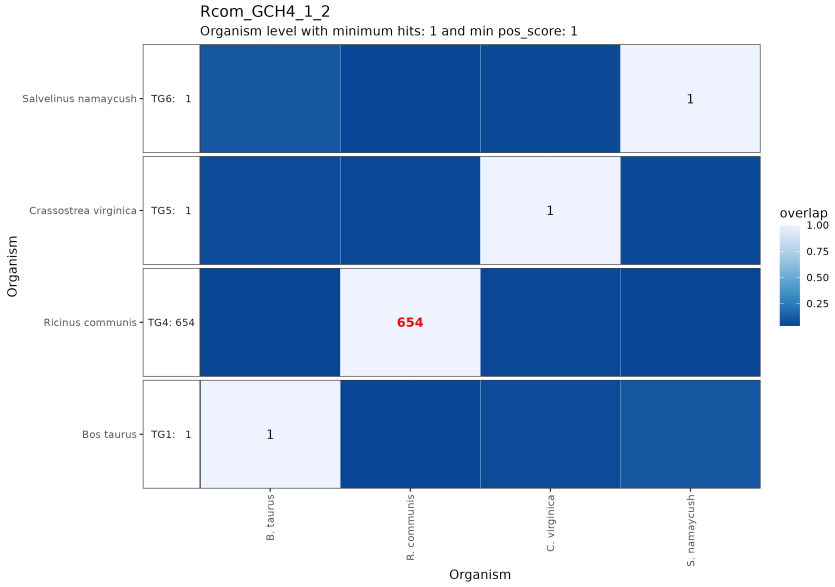

Figure 6: UniProt Heat Map R. communis (castor bean),Rcom_GC4 Sample

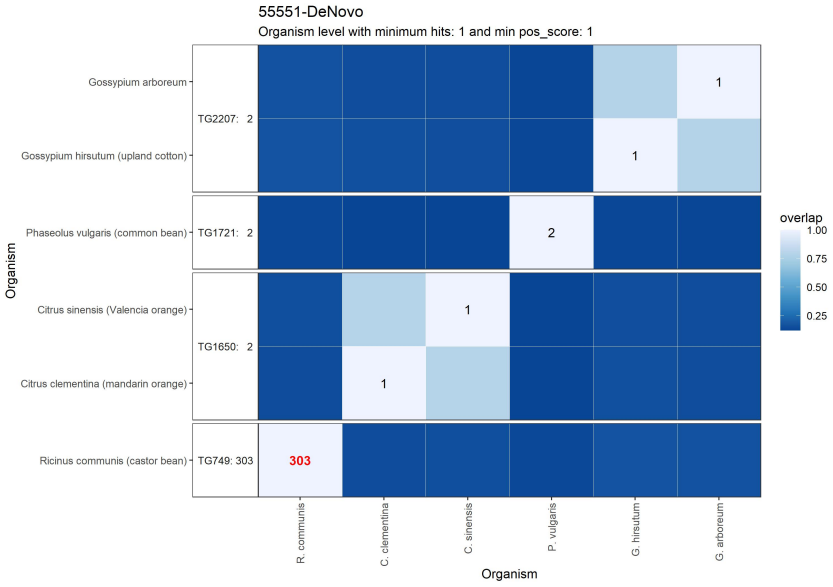

Figure 7: KEGG Heat Map R. communis (castor bean)

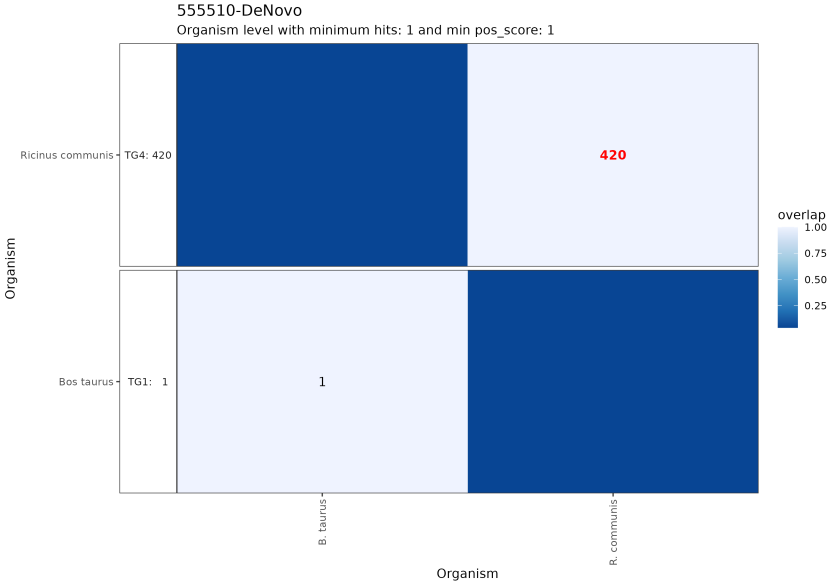

Figure 8: UniProt Heat Map R. communis (castor bean)

Ricin Prep Sample that was not found in KEGG or UniProt

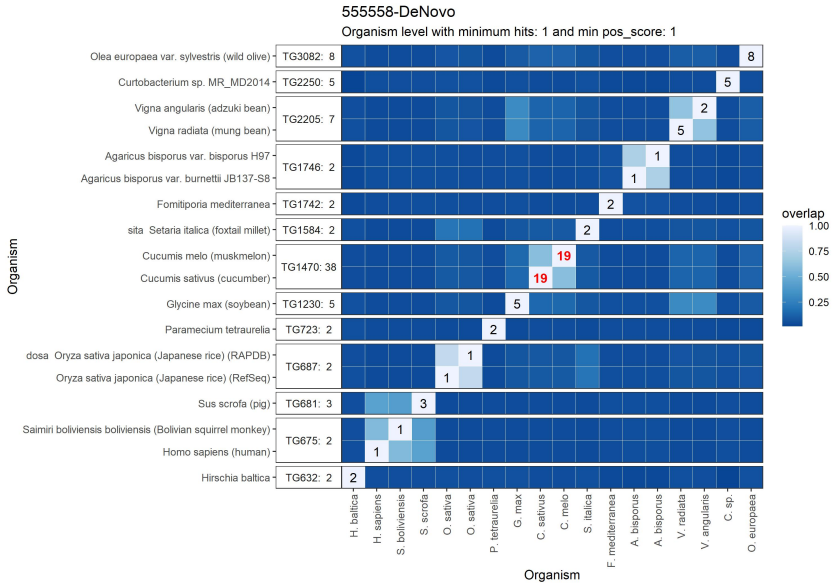

Figure 9: KEGG Heat Map, Ricin Prep that tested poorly on both KEGG and UniProt databases

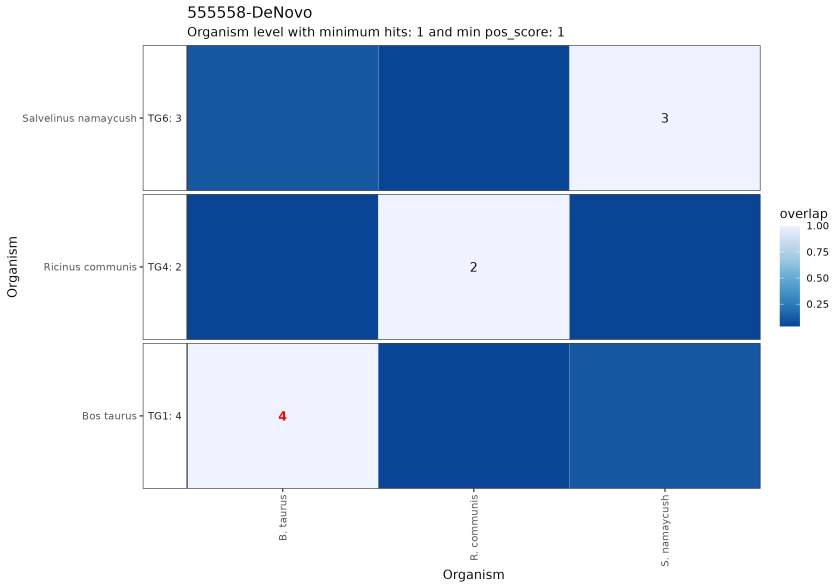

Figure 10: UniProt Heat Map R. communis Prep (castor bean)

Samples Tested which do not contain R. communis

Fish

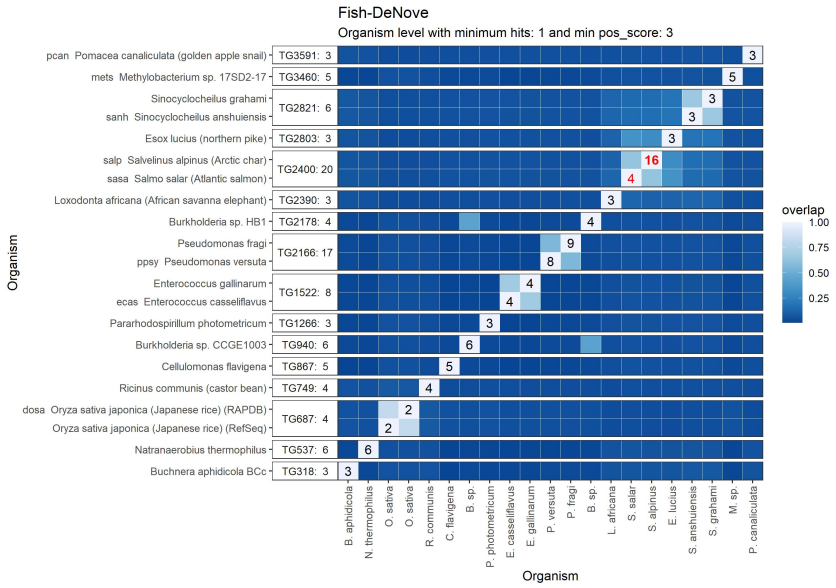

Figure 11: KEGG Heat Map S. namycush (lake trout)

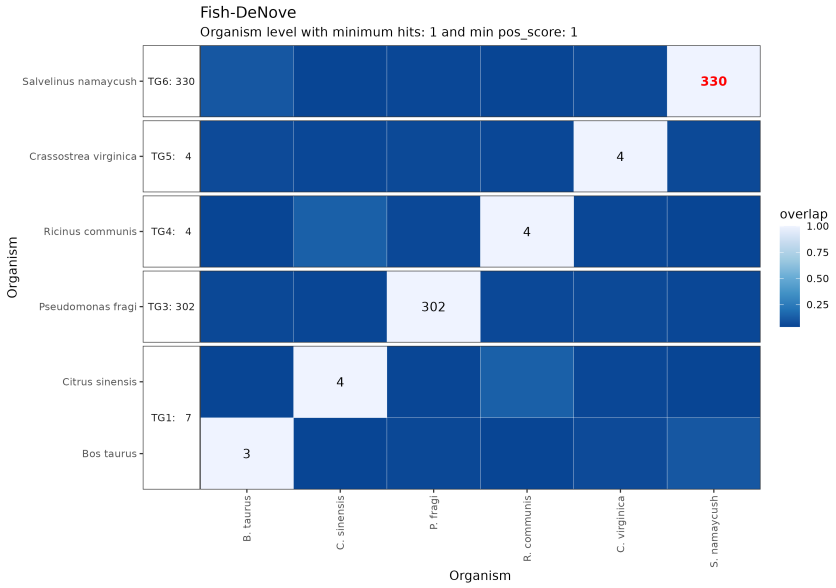

Figure 12: UniProt Heat Map S. namaycush (lake trout) with P. fragi (bacteria resulting from spoilage)

Orange Juice

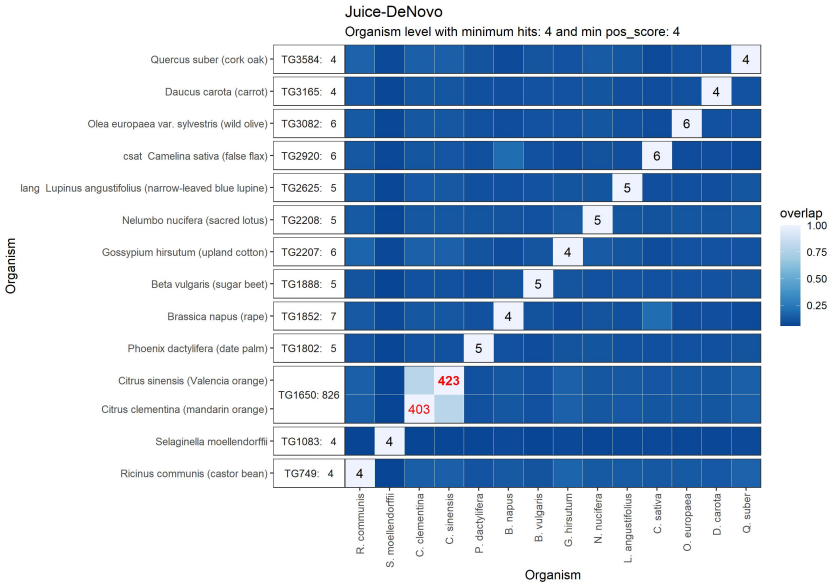

Figure 13: KEGG Heat Map C. clementina and C. sinensus (oranges)

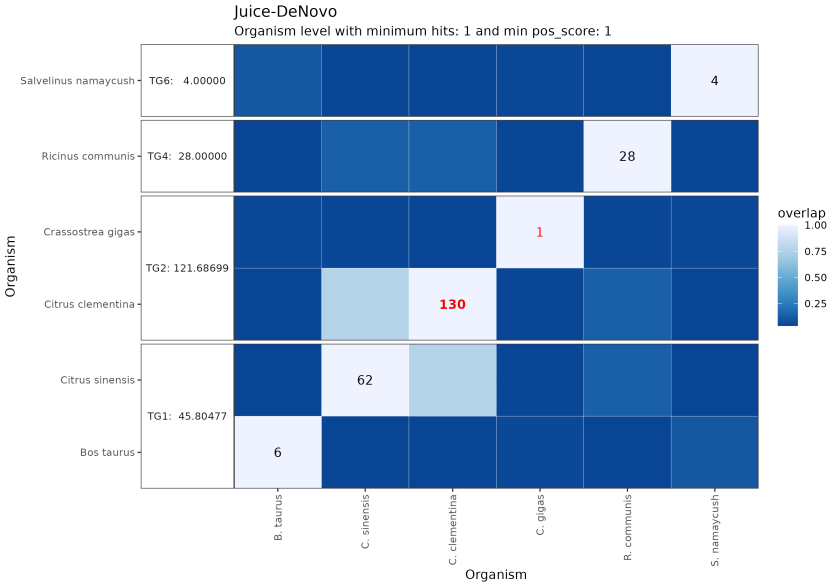

Figure 14: UniProt Heat Map C. clementina and C. sinensus (oranges)

Milk

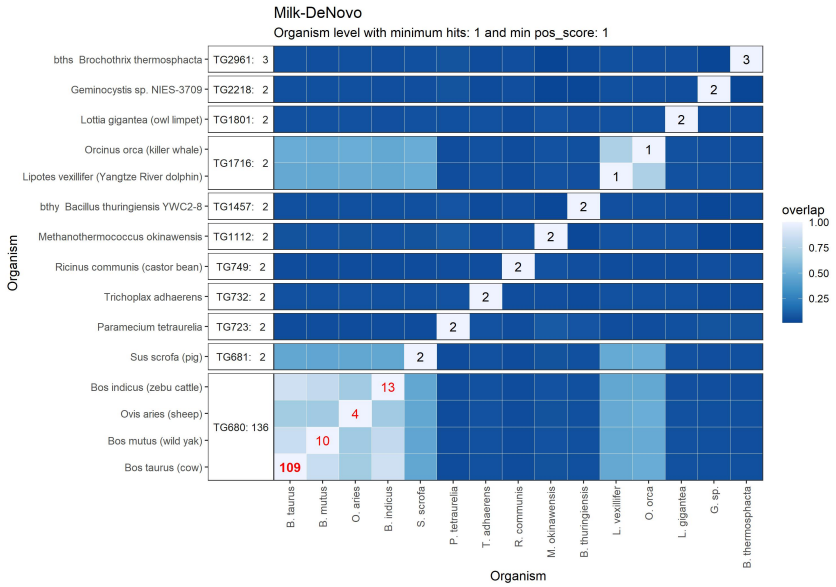

Figure 15: KEGG Heat Map B. taurus (milk)

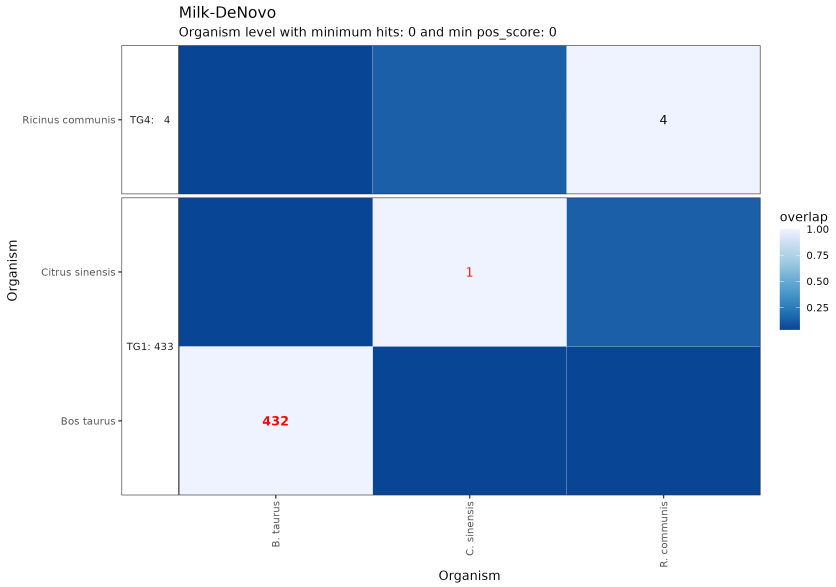

Figure 16: UniProt Heat Map B. taurus (milk)

Oyster

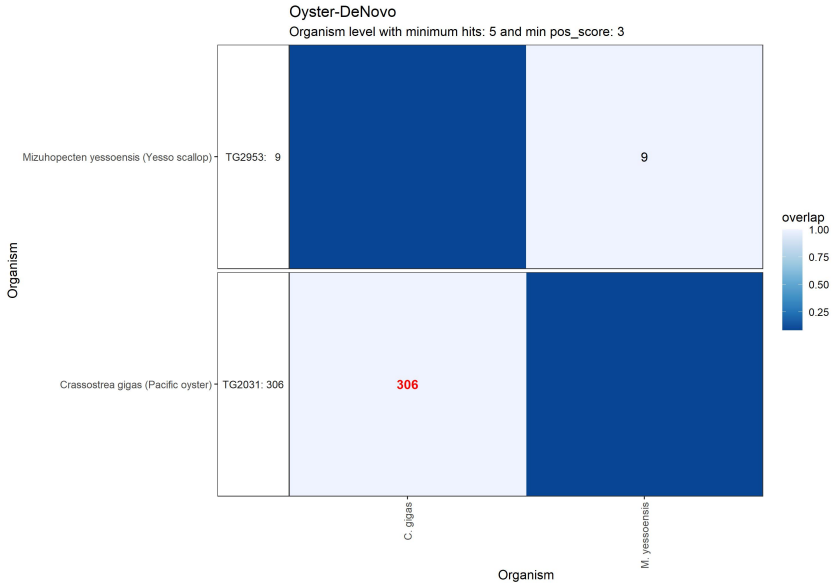

Figure 17: KEGG Heat Map C. virginica (oyster)

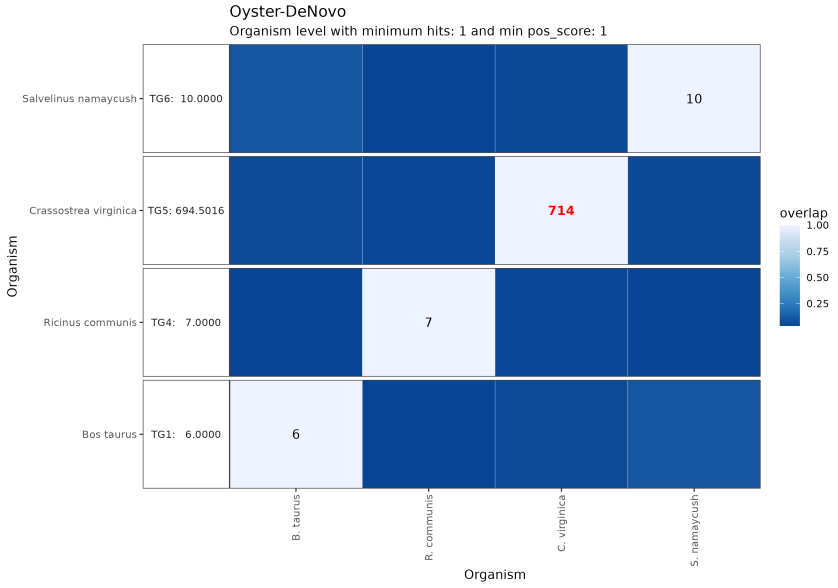

Figure 18: Uniprot Heat Map C. virginica (oyster)

References

Consortium, T. U. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Csordas, A., Ovelleiro, D., Wang, R., Foster, J. M., Ríos, D., Vizcaíno, J. A., & Hermjakob, H. (2012). PRIDE: Quality control in a proteomics data repository. *Database*, *2012*.

Dau, T., Bartolomucci, G., & Rappsilber, J. (2020). Proteomics using protease alternatives to trypsin benefits from sequential digestion with trypsin. *Analytical Chemistry*, *92*(14), 9523–9527.

Demirev, P. A., & Fenselau, C. (2008). Mass spectrometry in biodefense. *Journal of Mass Spectrometry*, *43*(11), 1441–1457.

Duracova, M., Klimentova, J., Fucikova, A., & Dresler, J. (2018). Proteomic methods of detection and quantification of protein toxins. *Toxins*, *10*(3), 99.

Jansen, H.-J., Breeveld, F. J., Stijnis, C., & Grobusch, M. P. (2014). Biological warfare, bioterrorism, and biocrime. *Clinical Microbiology and Infection*, *20*(6), 488–496.

Kanehisa, M.others. (2002). The KEGG database. *Novartis Foundation Symposium*, 91–100. Wiley Online Library.

Pal, M., Tsegaye, M., Girzaw, F., Bedada, H., Godishala, V., & Kandi, V. (2017). An overview on biological weapons and bioterrorism. *American Journal of Biomedical Research*, *5*(2), 24–34.

Pere, B. (2020). HOW CAN WE USE SHOTGUN PROTEOMICS TO IDENTIFY AND DISTINGUISH RCA60 FROM CLOSELY RELATED RICIN-LIKE PROTEINS AND PROTEINS FROM OTHER SPECIES?". *Capstone*.

Riedel, S. (2004). Biological warfare and bioterrorism: A historical review. *Baylor University Medical Center Proceedings*, *17*, 400–406. Taylor & Francis.

Sinha, A., & Mann, M. (2020). A beginner's guide to mass spectrometry–based proteomics. *The Biochemist*, *42*(5), 64–69.

Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C., & Yates III, J. R. (2013). Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews*, *113*(4), 2343–2394.

Zhao, Y., & Lin, Y.-H. (2010). Whole-cell protein identification using the concept of unique peptides. *Genomics, Proteomics & Bioinformatics*, *8*(1), 33–41.