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

07 November 2022

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. An automated software tool which can search for matches against a proteome database is useful for forensic analysis of samples and an effective countermeasure to Bioterrorist attacks. Here we create a database for this purpose using publicly available proteomes from the Uniprot.org database (Consortium, 2020).

Contents

Introduction

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 (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 law enforcement's forensic sample analysis. The accuracy of MS can also distinguish between similar compounds and organisms such as the castor bean vs the derrivative 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.

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 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 compares these peptides to a MySQL database which has been constructed from the Japanese 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. MARLOWE then uses an algorithm which compares the MS/PEAKS® data with the database peptides and makes a list of candidate organisms that may be contained in the sample. It produces scores and a heat map showing the most likely hits (Figure 1)

Specific Aims

Aim 1 is to create a new R package which will extract required fields from publicly available UniProt data [@10.1093/nar/gkaa1100] 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 site 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 not standard so these subroutines are not useful for any other 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 no longer required.

I created a new UniProt based peptide database by downloading data from Uniprot's uniref50 in fasta format and parsed it via the parse_fasta package to extract the data fields that MARLOWE requires. Creating the MARLOWE database from complete uniref50 data is a computationally intense process taking weeks on a powerful workstation or server. There are over 100,000 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 an NBACC High Performance Computing (HPC) cluster. I have ported the MARLOWE code and database from Windows to Linux to make this possible.

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

I am in the process of installing MARLOWE on a home server running Ubuntu 20.04LTS which has been upgraded with a high speed NVME SSD to store the database. I have also installed it on a Windows workstation at NBACC. MARLOWE was designed to run with Microsoft Windows® but I have converted it to run on Linux. I will test it on Ubuntu 20.04LTS and CentOS 7, and Red Hat RHEL 8 Linux.

I have loaded the MARLOWE code into a GitHub.com repo so I can fork from the current version and create new packages which can be tested on various linux platforms. The code is written with R version 4.2.1 using the RStudio Development Environment 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.

I will validate the output of MARLOWE (current version MySQL 8.0 on Windows 10, R version 3.6.1) with a few samples, including the Abrin sample, Milk, and Castor bean. To perform the analysis of the two databases, MARLOWE will be run against each database with the same samples. I will compare the output with the actual contents of the sample and make a conclusion on performance or improvements needed.

```
(base) danvogel@hpc0:~/R/marloweU/data-raw/uniprot$ grep ">" castor.head.fasta
>sp|O65039|CYSEP_RICCO Vignain OS=Ricinus communis OX=3988 GN=CYSEP PE=1 SV=1
>sp|P01089|2SS_RICCO 2S seed storage albumin protein OS=Ricinus communis OX=3988 PE=1 SV=2
>sp|P02879|RICI_RICCO Ricin OS=Ricinus communis OX=3988 PE=1 SV=1
>sp|P06750|AGGL_RICCO Agglutinin OS=Ricinus communis OX=3988 PE=1 SV=1
>sp|P22337|STAD_RICCO Stearoyl-[acyl-carrier-protein] 9-desaturase, chloroplastic OS=Ricinus communis OX=3988 PE=1 SV=1
>sp|Q41131|FAH12_RICCO Oleate hydroxylase FAH12 OS=Ricinus communis OX=3988 GN=FAH12 PE=1 SV=2
>sp|Q41142|PLDA1_RICCO Phospholipase D alpha 1 OS=Ricinus communis OX=3988 GN=PLD1 PE=1 SV=1
>sp|A1A442|DGAT2_RICCO Diacylglycerol O-acyltransferase 2 OS=Ricinus communis OX=3988 GN=DGAT2 PE=1 SV=1
>sp|B9RK42|GPC1_RICCO Glycerophosphocholine acyltransferase 1 OS=Ricinus communis OX=3988 GN=GPC1 PE=1 SV=1
>sp|B9RU15|ATXR5_RICCO Probable Histone-lysine N-methyltransferase ATXR5 OS=Ricinus communis OX=3988 GN=ATXR5 PE=1 SV=1
>tr|B9RUF4|B9RUF4_RICCO Peroxidase OS=Ricinus communis OX=3988 GN=RCOM_0852470 PE=3 SV=1
```

Figure 1: Fasta Communis Ricinis Sample

Results

Describe the results of the proposed project work.

Discussion

Putting the results in context

Summary and Future Work

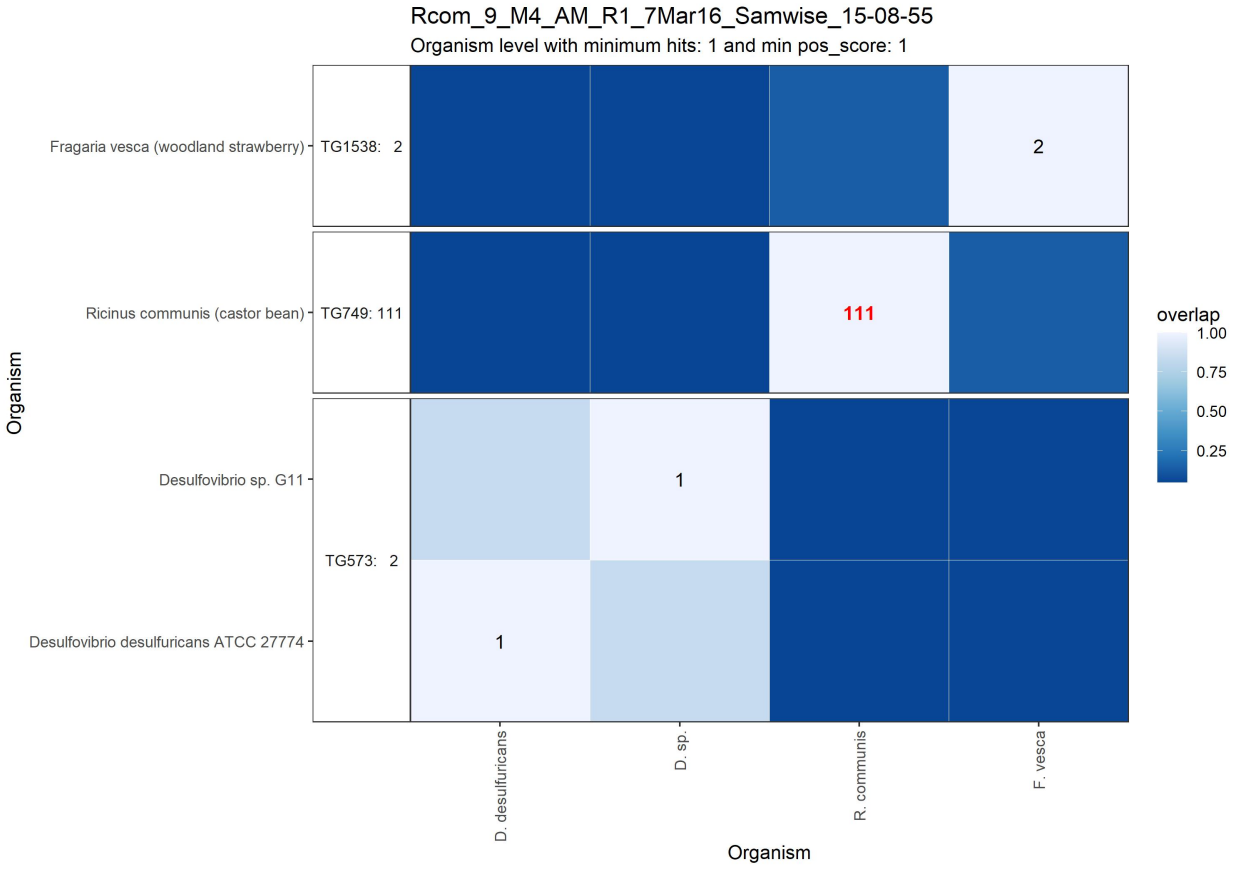
Alternate digestion enzyme

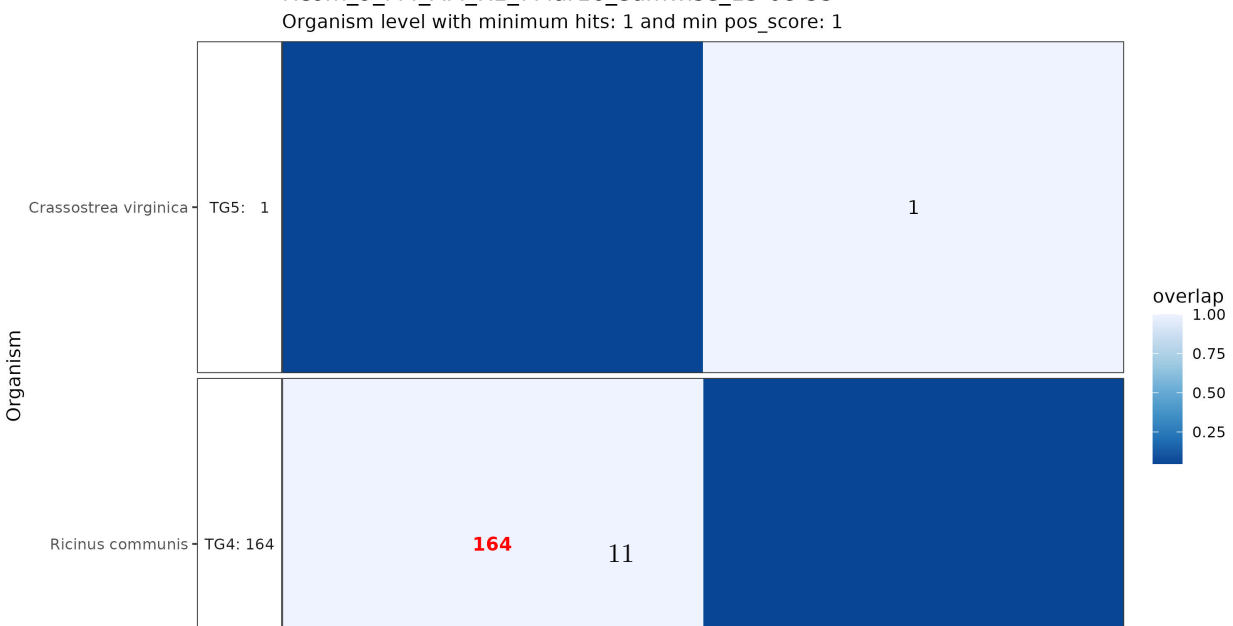
Currently MARLOWE only supports Trypsin digest. We can construct another version of the database where the peptides have been digested with an alternate protease enzyme.

Software 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


Rcom_9_M4_AM_R1_7Mar16_Samwise_15-08-55
Organism level with minimum hits: 1 and min pos_score: 1
Fragaria vesca (woodland strawberry)
Ricinus communis (castor bean)
Desulfovibrio sp. G11
Desulfovibrio desulfuricans ATCC 27774
TG1538: 2
TG749: 111
TG573: 2
2
111
1
1
Organism
D. desulfuricans
D. sp.
R. communis
F. vesca
Organism
overlap
1.00
0.75
0.50
0.25


Rcom_9_M4_AM_R1_7Mar16_Samwise_15-08-55
Organism level with minimum hits: 1 and min pos_score: 1
Crassostrea virginica
Ricinus communis
TG5: 1
TG4: 164
1
164
11
Organism
overlap
1.00
0.75
0.50
0.25

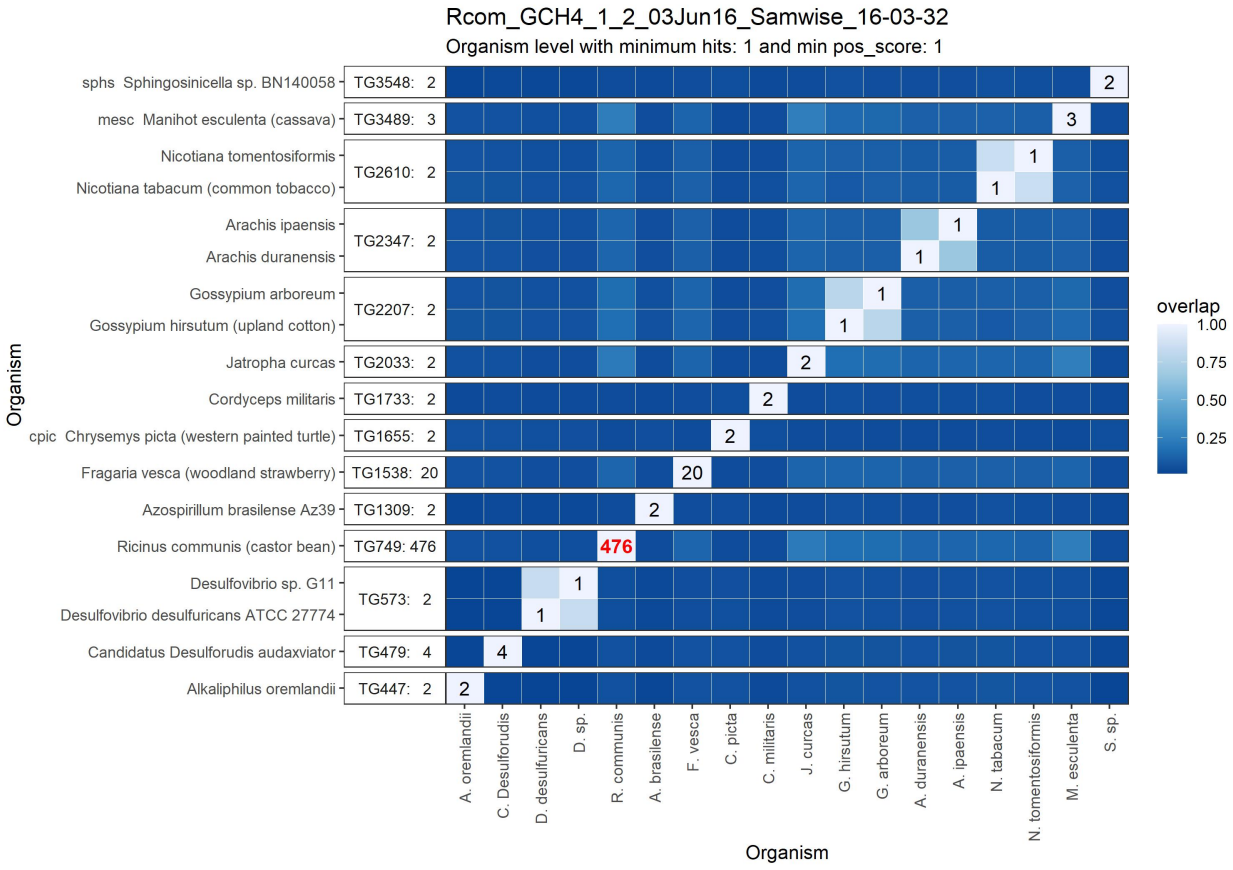


Figure 2: Heat Map Communis Ricinis Rcom_GC4 Sample on KEGG

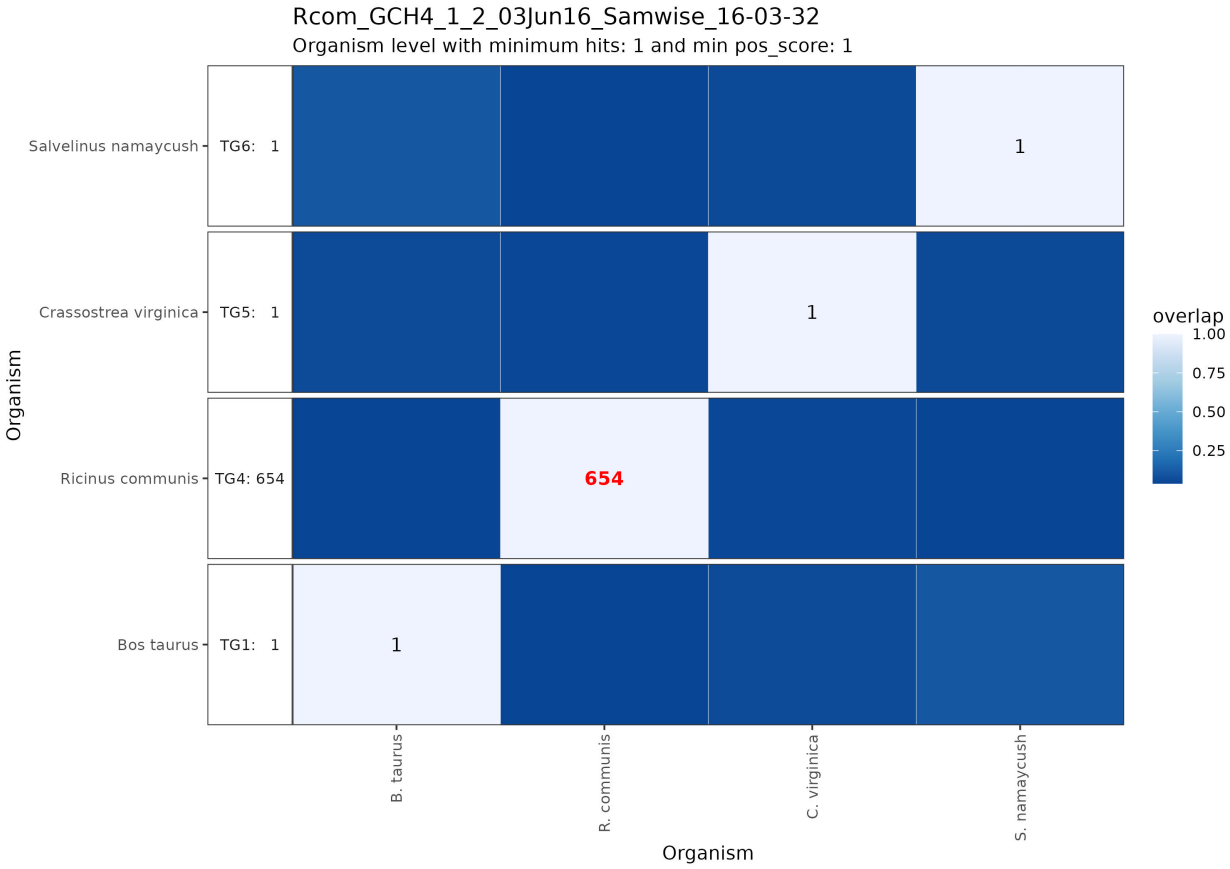


Figure 3: Heat Map Communis Ricinis Rcom_GC4 Sample on UniProt

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.