Universiteit Leiden
The Netherlands

Opleiding Informatica

The Intestinal Flora of the Zebrafish

Per Hermanus & Sjoerd Wesselman

Supervisors:
M.N. Palmblad & H.P. Spaink & F.J. Verbeek

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

16/08/2019

Abstract

The zebrafish is a model organism necessary in drug development. Therefore it is important to have a good understanding of its proteome. The presence of a microbiome in its intestine provides a challenge when researching the proteome. In this study we attempt to differentiate the proteome of the zebrafish from the proteome of its microbiome. Using the Trans-Proteomic Pipeline and LC-MS/MS data from the muscle and intestine of the zebrafish we created a workflow to show peptide and protein distribution in both tissues. Our results are not able to demonstrate a significant distinction in distribution of peptides and proteins to be able to conclude that there is a microbiome present.

Contents

1 Introduction

The zebrafish (Danio Rerio) is an important organism in scientific research. They are widely used as a model organism. Model organisms are an important part in biomedical research, such as finding mechanics of diseases and drug development. To be able to do this research it is important to have knowledge of DNA, mRNA, proteins and metabolites. Studies into these fields are respectively called, genomics, transcriptomics, proteomics and metabolomics. All fields need to be combined when using animal models. This study will focus mainly on proteomics to find out more about zebrafish. [vdPD]

1.0.1 Proteomics

Proteomics is the field of study where proteins and their interactions are studied. When DNA is read, messenger RNA is created. This mRNA in turn can be read by the ribosomes in a cell to create proteins. The existence and concentration of a protein can hugely affect the cell and, if it is part of one, the multi-cellular structure it resides in. The composition of proteins at a given time is called the proteome. The proteome can tell us a lot about the condition of an organism since the proteome of a cell or organism changes based on its needs and the influences it experiences [AJG+16]. In our case we use proteomics to explore the makeup of certain tissues of the zebrafish, in particular the intestines and its flora.

The basis of large-scale protein analysis, or proteomics, rests upon Frederick Sangers work on DNA, RNA and protein sequencing. Since its conception in the latter half of the 20th century, many advancements have been made in the speed and scale of sequencing. Usability improvements such as automated sequencing [EB67], automated data collection/analysis [Hun81] and the decrease in sample material needed followed. This allowed for the creation of large sequence databases.
Which proteins in a cell occur in what quantity is called the proteome of the cell. The proteome can tell a lot about the state of the cell at a given moment. Protein extraction and separation trough 2D gel electrophoresis allowed scientists to get a rough idea of the proteome by looking at the stained gels. Currently, we can better analyse the results of these gels trough mass spectrometry and computer programs, creating a list of proteins. This gives researchers the ability to look up the names of found proteins, their functions, their corresponding genes and their origin. [JAM97]

1.0.2 Zebrafish

There are multiple reasons why zebrafish are widely used as model animals. They are easy to keep and manage in a laboratory and they produce around 200 eggs per week, which are naturally transparent. This makes ideal for microscopic imaging. They develop fast, especially in the early stages of their life. They are also relatively closely related to humans. Around 70% of their protein encoding genes are similar to human genes. [HCT+13].

1.0.3 Microbiome

In this study will focus on the intestine of the zebrafish. Usually, there is a microbiome present in the intestine of an organism. A microbiome is “the ecological community of commensal, symbiotic, and pathogenic microorganisms” [JM01] in a multicellular organism.

1.1 Problem

We want to explore the proteome of the zebrafish. This has been done in previous research for most of the tissues of the zebrafish [vdPDMD+14]. The intestines, however, are a challenge. Using hierarchical clustering of the data from each organ it has been shown that the intestines differ from all other organs (see Figure 1). We suspect that the presence of a microbiome in the intestines may play a role. Therefore it is needed to find a way to identify the proteins of the microbiome and the proteins of the zebrafish.

Figure 1: The hierarchical cluster of tissues in zebrafish based on LC-MS/MS data. [vdPDMD+14]

1.2 Research Questions

Our main research question for this thesis will be: Can we demonstrate the presence of a microbiome in the zebrafish its intestine and can we differentiate the proteome of zebrafish from the proteome of the microbes in its intestine?
To answer this question the following sub-questions need to be answered.

1. What does the proteome of zebrafish usually look like?
2. Which microbes are expected to be present in the intestines of the zebrafish?
3. Is it possible to identify all proteins found in the intestine of the zebrafish?
4. What percentage of intestinal proteins comes from the zebrafish itself and what comes from its intestinal flora?

1.3 Thesis structure

This chapter contains the introduction. Section 2 gives an overview of all methods and materials used during this research. In section 3 we give a detailed description of all actions performed to do this research. Section 4 shows the results from our research. In section 5 we give a conclusion and discuss the results.

1.4 Contributions

This thesis was supervised by Dr. N.M. Palmblad of the LUMC, Prof. dr. ir. F.J. Verbeek of the LIACS and Prof. dr. H.P. Spaink of the IBL.
S. Wesselman mainly focused on creating the sequence databases and pre-processing the data. While P. Hermanus focused more on testing and deciding on paramaters to use and the post-processing of the data. Finding the results and making important decisions was all done together.

2 Methods and Materials

2.1 Microbiomes

From researchers of the Institute of Biology Leiden, we received a list of seven bacteria species that are expected to be present in the intestine. Five of these bacteria have been discovered in previous research. [RMS+11] The following bacteria strains are used for research by the Biology Institute.

- Aeromonas ZOR0001 (Proteobacteria)
- Plesiomonas ZOR0011 (Proteobacteria)
- Vibrio ZWU0020 (Proteobacteria)
- Shewanella ZOR0012 (Proteobacteria)
- Pseudomonas ZWU0006 (Proteobacteria)
- Exiguobacterium ZWU0009 (Firmicutes)
- Chryseobacterium ZOR0023 (Bacteroidetes)

The genomes of Chryseobacterium, Exiguobacterium, Pseudomonas and Shewanella were available to us. The proteome of Exiguobacterium was already available on UniProt as a reference proteome.

2.2 Datasets

For this project, we will be working on datasets of the zebrafish proteome created by S. van der Plas-Duivesteijn, provided to us by N.M. Palmblad. S. Arampatzi provided four genome assemblies to us. Several FASTA databases have also been created with sequences found on UniProt and NCBI.

2.2.1 Raw Data

The datasets we will be working with contain LC-MS/MS data from multiple tissues. This data is already converted to the mzXML format. Apart from intestinal data, we'll also work with data from muscle tissue. While there was also data from the skin tissue available to us, we were unable to use this data due to time constraints.
A number of steps have been taken in the creation of the datasets we used. First, two separate protein extractions were performed on each tissue. These extracted proteins were then separated with the use of SDS-Page gel separation. Cutting these gels yielded slices with proteins which have been digested in-gel, in preparation for LC-MS/MS (Liquid Chromatography Tandem Mass Spectrometry). After each MS scan, a maximum of 10 of the most frequently occurring multiply charged ions in the 300-1300 m/z range were selected for MS/MS [vdPDMD+14]. The resulting raw mass spectrometry data was then converted into the mzXML files our dataset consists of. Each mzXML file corresponds to a slice from the SDS-Page gel.

2.3 Sequence database

To be able to find which proteins are present in the intestine we would need a sequence database to search against. Since there was no database with protein sequences from both zebrafish and their intestinal microbiome available to us, one was created for this project.
We have created multiple FASTA files containing protein sequences. These files consist of two main parts. The proteome of the zebrafish and the proteomes of seven species of bacteria which were expected to be present in the intestine. For the zebrafish, we used the reference proteome from UniProt. For the bacteria, we used several different datasets. One option was to use GeneMarkS-2 [LGT+18] to convert the genomes available to us to predicted proteins. There is a reference sequence available for Exiguobacterium. We also used sequences available in the NCBI protein database using a taxonomy search. On Uniprot we were able to find all the proteins associated with the genus of each bacteria. Table 1 shows for each possible part of the database the number of proteins this dataset contained. Using this table we tried to create a database with an as equal as possible number of proteins for each bacterial strain.

Bacterial strains	Genome	NCBI	Uniprot Rev.	Uniprot Full
Aeromonas ZOR0001	–	3864	1012	292816
Chryseobacterium ZOR0023	12559	–	1	599779
Exiguobacterium ZWU0009	3502	3196	510	68471
Plesiomonas ZOR0011	–	3278	2	9338
Pseudomonas ZWU0006	10176	4659	462	22743
Shewanella ZOR0012	12086	4168	8464	261783
Vibrio ZWU0020	–	3672	5248	1307465

Table 1: The amount of proteins available for each strain for each different search method used. The genome column shows the number of proteins found using GeneMarkS-2 on the genomes. The NCBI column shows the number of proteins found using a NCBI taxonomy search. The Uniprot Rev. column shows the number reviewed proteins found on Uniprot, whereas the Uniprot Full shows the number of reviewed and unreviewed proteins.

2.4 Equipment

A large number of specialized equipment and machinery was used during the extraction process and the creation of the raw data. For this thesis, the most important aspect is the setup used for LC-MS/MS. The LC-MS/MS system consists of a NanoLC-Ultra 2D plus controlled by HyStar 3.2 coupled to an amaZon speed ETD ion-trap with an Apollo II ElectroSpray Ionization source controlled by trapControl 7.1 [vdPDMD+14].
No specialised machinery was used in either analysis of the raw data or the creation of the FASTA databases.

2.5 Software and Hardware

The windows Petunia GUI version of the trans-proteomic pipeline (TPP) [DMS+15] was used to analyse the files. The following tools from the TPP are used:

- Comet Search
- PeptideProphet
- ProteinProphet

We have also written several small python scripts and used GeneMarkS-2 to help to create the sequence database and to analyze the data. Two systems have been used to run the TPP.

- A server running Windows 10 Pro for workstations with Dual Intel Xeon E5645@2.4GHz hexa-core 12 thread processors and 20GB of RAM.
- A PC running Windows 10 Pro with an Intel Core I7 6700k@4.0GHz quad-core 8 thread processor and 16GB of RAM

3 Design and Implementation

3.1 GeneMarkS-2

GeneMarkS-2 [LGT+18] is a model that can identify genes when given a genome from a prokaryote. We used it to convert the genomes that were available to us to genes and convert those to protein sequences.

3.2 Comet

Comet is an open-source tandem mass spectrometry sequence database search tool. It searches the raw mass spectra against a provided sequence database for peptides. It does this by implementing a cross-correlation algorithm which scores peptide sequences against experimental mass spectra, generating an Expect-value. [EJH13]
The parameters needed highly depend on the MS method used. During our initial testing, we used the "high-low" settings (2018.01 rev. 4) found on the Comet website [Com]. This produced very inaccurate results since the MS/MS was done via ion-trap 2.4. Using the "low-low" parameter file (2018.01 rev. 0) meant for "low res MS1 and low res MS2 e.g. ion trap" resulted in much more reliable results. As such all data was searched using the default "low-low" parameters.

3.3 PeptideProphet

PeptideProphet [KNKA02] is an algorithm that can validate peptide assignments made by algorithms like Comet. It estimates the accuracy of the assignments made by Comet to the tandem mass spectra. It uses the distribution of search scores and peptide properties from correct and incorrect peptides to calculate the probability that a given peptide is assigned correctly.

3.4 ProteinProphet

After analysing with PeptideProphet the data is usually further analysed with ProteinProphet [NKKA03]. ProteinProphet validates protein assignments based on the peptides assigned by Comet and PeptideProphet. ProteinProphet gives a probability to each protein that has been predicted. This probability tries to deal with two issues in protein prediction:

1. Proteins that have only one corresponding peptide ('single-hit proteins') are less likely to be correct than proteins with multiple corresponding peptides ('multi-hit proteins').

2. Peptides can be assigned to multiple proteins in the sequence database. ProteinProphet tries to find the simplest list of proteins that can describe the detected peptides (Occam's Razor).

Proteins that can be explained by the same peptides in ProteinProphet are called a group. ProteinProphet uses Occam's Razor and gives the proteins that are not part of the simplest solution a probability of 0. We have chosen to leave these proteins out of our results because they are most likely false positives [Pro].

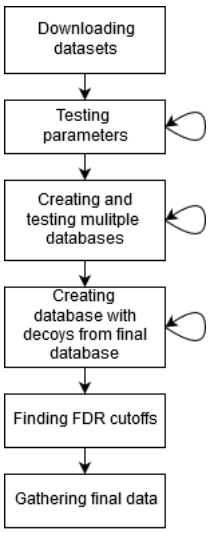

Figure 2: All the steps taken during our research to create the final workflow and our results

3.5 Implementation

We had four different MS/MS databases at our disposal: the intestine from both a male and female zebrafish and the muscle from both a male and female zebrafish. The muscle database was used as null test. The muscle dataset was chosen, because it should contain almost no bacterial data.

To start, we tried multiple combinations of parameters for both Comet and PeptideProphet on the female intestine dataset using the zebrafish reference sequence as search database (Figure 2). After testing multiple parameters we decided to use the Comet low-low parameters and PeptideProphet with default settings. These settings resulted in a PeptideProphet model that seemed sufficient, whereas other settings resulted in unexpected results. We mainly used the error/sensitivity plots and the histograms of search results provided by PeptideProphet to determine if a model was sufficient. A model is ideal when the area between the sensitivity and error is as large as possible. The histograms should show two peaks. One peak represents the bad hits, which are modelled by the red curve. The other peak represents the good hits, which are modelled by the green curve. If there is a clear separation between the positive and negative distributions, you can say the model is sufficient.

We created four different sequence databases to search against. One for each of the columns in table 1. If there did not exist a dataset for a given bacterial strain in that column, we used the Uniprot Reviewed dataset for that bacterial strain instead. The first three sequence datasets gave comparable results. We were unable to run the Full Uniprot dataset because of the high amount of computing time needed. Therefore we decided to find the best mix of the first three databases. This was done in such a way that all seven bacterial strains had a comparable amount of proteins in the final database. This database consisted of the NCBI datasets of all bacteria except chryseobacterium. For chryseobacterium we used the database created by GenemarkS-2 using the genome data. The zebrafish dataset was the UniProt reference dataset. We tested all four MS/MS datasets against this database. We used the TPP together with a pre- and post-processing scripts written in python

to process the data (Figure 3).

Then we created a decoy database to find the best cutoff points (the minimum probability a peptide should be assigned to be taken into account). Our decoy database is a concatenation of the database we were trying to find the cutoff point for and a randomised version of that database. This database was created using the "decoy databases" tool in TPP. We then used Comet to search this database for peptides and used PeptideProphet with a non-parametric model. Using this we were able to determine the cutoff point where the false discovery rate (FDR) equals 1%. The cutoff points chosen are shown in table 2.

Dataset	Cutoff
Intestine male	0.920
Intestine female	0.927
Muscle male	0.920
Muscle female	0.905

Table 2: The cutoff points calculated for each of the data sets

3.6 Scripts

We wrote a script to combine the different FASTA files into one large FASTA file. This copied all the downloaded FASTA files into the final file. When a file was created by GeneMarkS-2 it would alter the file by adding the organism name and id to each protein sequence.
Two scripts were used to obtain the number of times an organisms peptide or protein was found in the database. These were run on TSV files exported trough the pepXML and protXML viewers. First, a script that makes sure the protein column contains the name of the species of origin. Some proteins only had an accession number in the protein column. These were all the proteins where the accession number started with "WP". We wrote a script that would take the name of the protein from the description column and add it to the protein column.

```
tsvfile = tsv file exported from pepxml or protxml data
database = orginal FASTA file used in Comet search
output = file output will be written to as tsv

for each row in tsvfile:
    proteins = all proteins in protein column on this row
    newstring = empty string
    for each protein in proteins:
        if proteins starts with ‘‘WP’’:
            species = string between ‘[’ and ‘]’ in description
            newstring += protein + ‘|’ + species + ‘,’
            break
        else
            newstring += protein + ‘,’
```

```
        tsvfile[currentrow][protein] = newstring

    output = tsvfile
```

This is followed by a script that checks for certain keywords in the protein column, allowing the creation of a list with the number of occurrences for each species. Sometimes a protein can be found in multiple species. In this case $1/n$ with n being the number of species will be added to each found species unless it can originate from the zebrafish, then 1 will be added to zebrafish.

```
for file in Directory:
    import file
    fileframe = pandas.readcsv(file)
    Initialise list with expected species
    for each row in fileframe:
        for each expected species:
            check protein column for key-substrings
            #meaning the pep/prot in that row can
            #originate from this species
        if 1 specie is found:
            add 1 to species
        if n multiple species are found:
            if zebrafish is one of them:
                add 1 to zebrafish
            else:
                add 1/n to each found specie

        if unknown specie is found:
            add to specie list and add 1
    import lists in pandas dataframe
    export dataframe as CSV
```

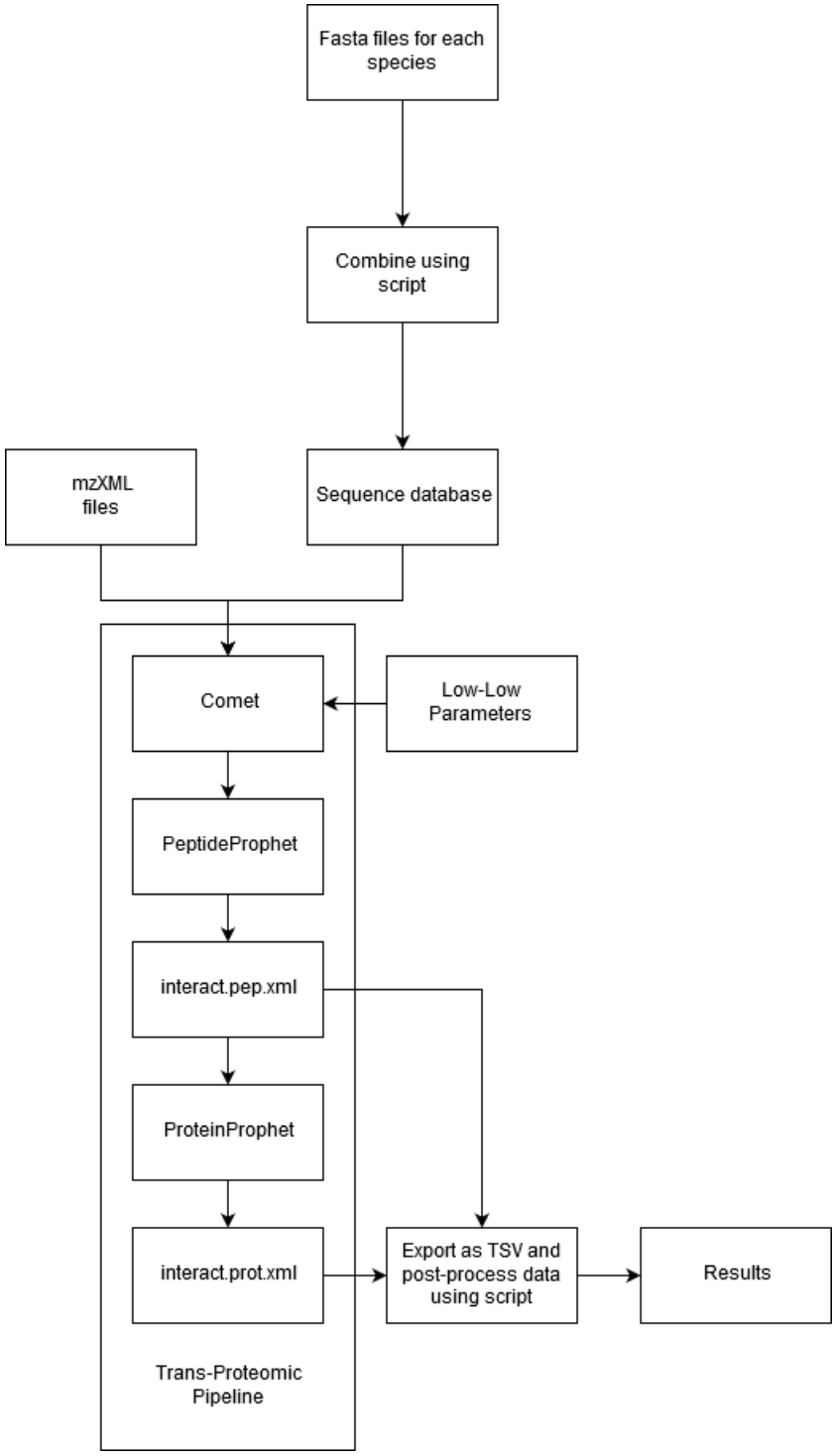

Figure 3: The flow of data trough the TPP including pre- and post-processing scripts.

4 Results

Using the implementation shown in Section 3 we generated the following results:

4.1 Peptides

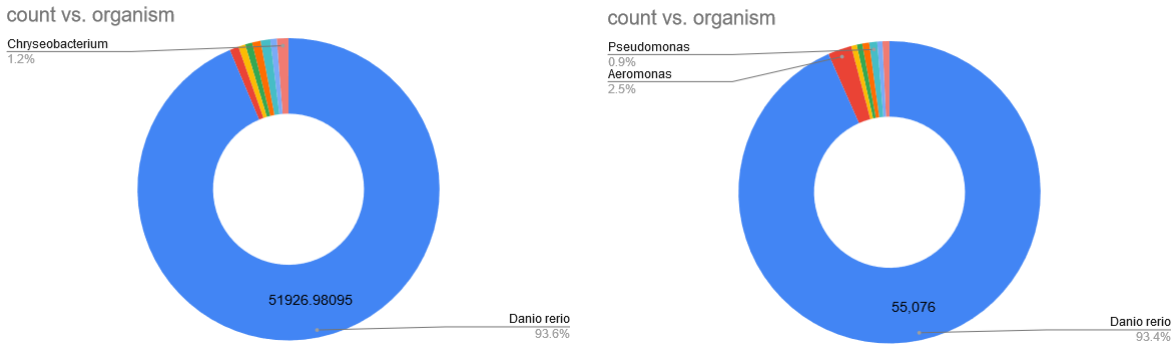

Figure 4: Male peptide distributions. Left:Muscle, Right:Intestine

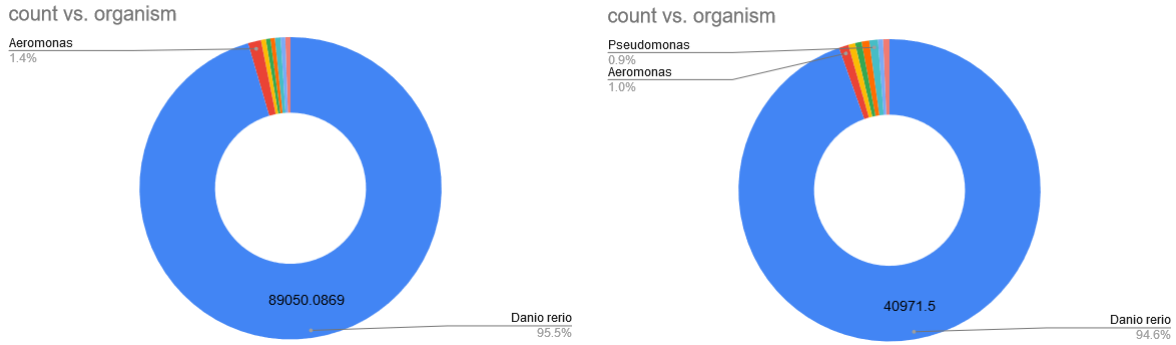

Figure 5: Female peptide distributions. Left:Muscle, Right:Intestine

The peptide distributions for both male (Figure 4) and female (Figure 5) show that the majority of peptides found originates from the zebrafish. This is visible in the muscle and the intestine datasets of both genders. To verify our findings we used Unipept [MDA+15]. This created the sunburst graphs in figure 6 and 7.

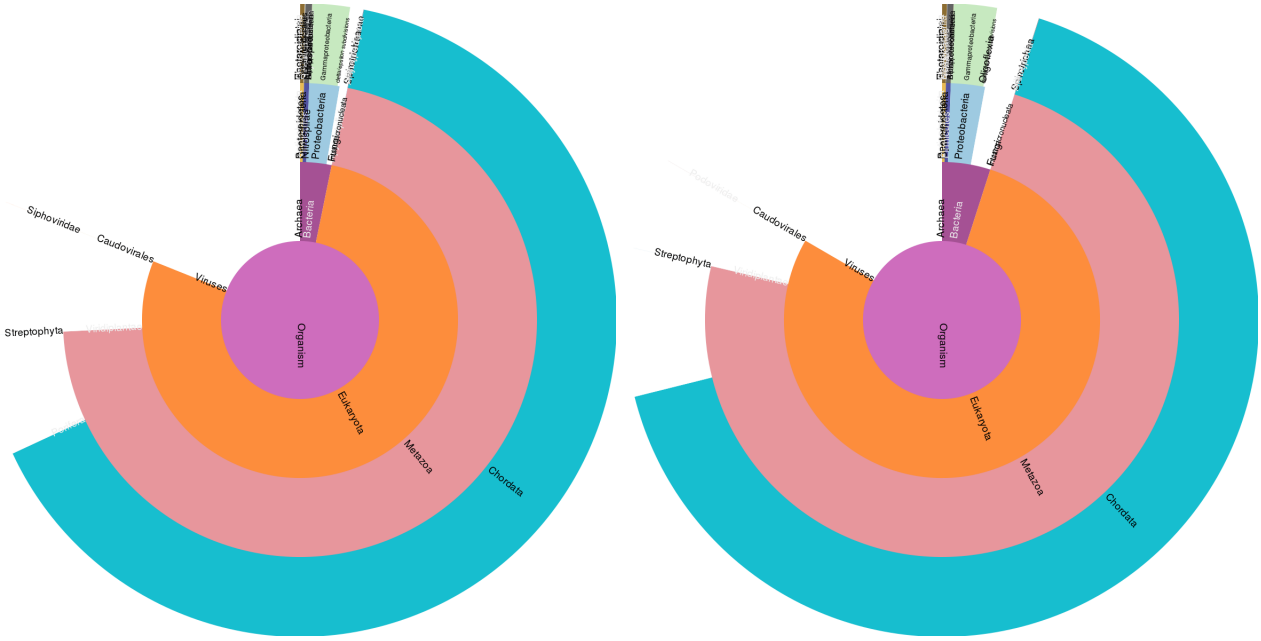

Figure 6: Male peptide distributions according to Unipept. Left: Muscle(3.2% bacteria), Right:Intestine(4.9% bacteria)

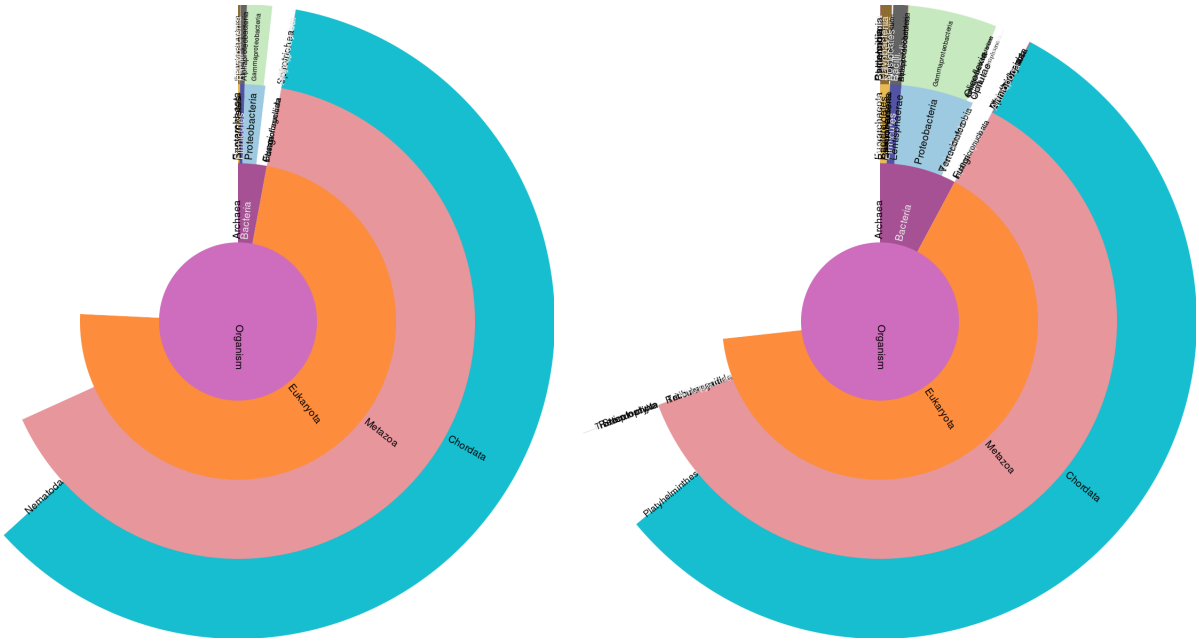

Figure 7: Female peptide distributions according to Unipept. Left:Muscle(4.4% bacteria), Right:Intestine(7.8% bacteria)

	muscle m	intestine m	muscle f	intestine f
Danio rerio	51926.98095	55076	89050.0869	40971.5
Aeromonas	536.547619	1497	1303.436905	449.3
Plesiomonas	420.997619	351	488.6702381	315.2666667
Vibrio	412.8666667	352	448.6583333	301.2666667
Shewanella	503.8309524	436	445.3369048	356.5166667
Pseudomonas	578.8142857	505	569.8869048	401.0166667
Exiguobacterium	380.8809524	324	461.3869048	234.5833333
Chryseobacterium	689.0809524	438	518.5369048	296.55
Other	5	8	3	2
Total	555455	58987	93289	43328

Table 3: Results of PeptideProphet analysis. Cut-off points can be found in table 2

4.2 Proteins

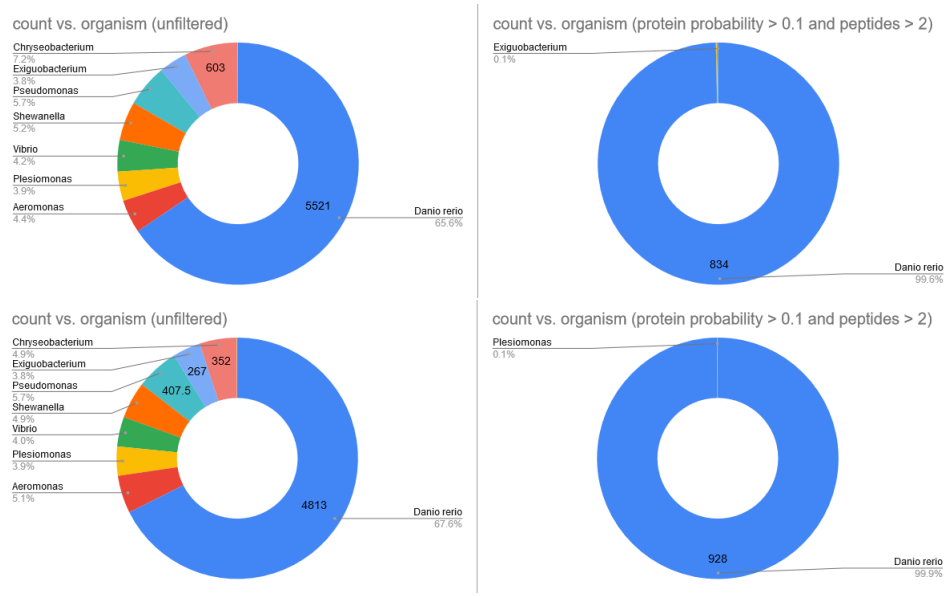

Figure 8: Male protein distributions. tl: muscle unfiltered, tr: muscle filtered, bl: intestine unfiltered, br: intestine filtered.

Comparing the unfiltered left charts with the peptide charts we can see that the distribution has shifted towards a larger percentage of micro-organism residue found. The right charts correct for ProteinProphet's "Occam's Razor" approach by cutting of near-zero percent protein probabilities and for unreliable data by taking only proteins with 3 or more found peptides into account (see Section 3.4) This removes nearly all bacterial findings, as can also be seen when comparing Table 4 against Table 5.

	muscle m	intestine m	muscle f	intestine f
Danio rerio	5521	4813	3905	4265
Aeromonas	374.1166667	360.0833333	302.5333333	281.5
Plesiomonas	329.2833333	277.9166667	220.8666667	225.8333333
Vibrio	354.7833333	283.0833333	284.8666667	256.3333333
Shewanella	437.2833333	351.4166667	304.7833333	303.1666667
Pseudomonas	478.5333333	407.5	363.45	331.1666667
Exiguobacterium	316	267	220	210
Chryseobacterium	603	352	319	244
Other	5	8	3.5	2
Total	8419	7119	5924	6119

Table 4: Unfiltered Proteinprophet results

	muscle m	intestine m	muscle f	intestine f
Danio rerio	834	928	582	839
Aeromonas	0	0	1	0
Plesiomonas	1	1	1	1
Vibrio	0	0	0	0
Shewanella	0	0	0	1
Pseudomonas	1	0	0	0
Exiguobacterium	1	0	0	0
Chryseobacterium	0	0	0	0
Other	0	0	0	0
Total	837	929	584	6119

Table 5: Proteinprophet results filtered on protein probability and number of peptides. This is done to remove groups and unreliable proteins from the data.

5 Conclusion and Discussion

Given our results, we can not conclude that there is a microbiome present in the intestines of the zebrafish. The results do not show a significant difference between the distribution of peptides and proteins in the intestine and distribution in muscle. It is expected that there is very little bacterial peptides in muscle. The bacterial peptides shown in muscle are therefore likely to be false positives. Because the intestine has a similar distribution it can't be determined if the bacterial peptides found in the intestine are true positives. Furthermore, most results show that only a small part of all peptides originate from the microbiome. The share of proteins originating from bacteria seems much larger at first. Most of the bacterial proteins, however, only have one peptide with a very low coverage. Given our large dataset it is highly likely these results are false positives. When correcting for these false positives it gives a result comparable with the peptides. Using Unipept to find results resulted in mostly the same distribution. This reinforces our conclusion.

5.1 Future Work

An explanation for the results can be that the intestines might be washed before the proteins were extracted. We can't say this with certainty. If this is the case then this study will need to be repeated with data from an intestine that has not been washed.

It could also be possible that there not enough data available on the bacteria suspected to be in the intestine to be able to find it. When there is more data available this study could be repeated.

Another possibility is that our approach was not the right one for this problem. Possibly, different parameters should be used or even a completely different pipeline can be used.

Acknowledgements

The authors thank Magnus Palmblad for his help and insight during our research, Fons Verbeek for his help on the paper and planning. We also want to thank Herman Spain for supervising and Suzanne Duijvestein and Semina Arampatzi for the data.

References

[AJG+16] Johnathon D. Anderson, Henrik J. Johansson, Calvin S. Graham, Mattias Vesterlund, Missy T. Pham, Charles S. Bramlett, Elizabeth N. Montgomery, Matt S. Mellema, Renee L. Bardini, Zelenia Contreras, Madeline Hoon, Gerhard Bauer, Kyle D. Fink, Brian Fury, Kyle J. Hendrix, Frederic Chedin, Samir EL-Andaloussi, Billie Hwang, Michael S. Mulligan, Janne Lehti, and Jan A. Nolta. Comprehensive proteomic analysis of mesenchymal stem cell exosomes reveals modulation of angiogenesis via nuclear factor-kappab signaling. *STEM CELLS*, 34(3):601–613, 2016.

[Com] Software:proteinprophet. http://comet-ms.sourceforge.net/. Accessed 16-08-2019.

[DMS+15] Eric W. Deutsch, Luis Mendoza, David Shteynberg, Joseph Slagel, Zhi Sun, and Robert L. Moritz. Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *PROTEOMICS Clinical Applications*, 9(7-8):745–754, 2015.

[EB67] P. Edman and G. Begg. A protein sequenator. *European Journal of Biochemistry*, 1(1):80–91, 1967.

[EJH13] Jimmy K. Eng, Tahmina A. Jahan, and Michael R. Hoopmann. Comet: An open-source ms/ms sequence database search tool. *PROTEOMICS*, 13(1):22–24, 2013.

[HCT+13] Kerstin Howe, Matthew Clark, Carlos Torroja, James Torrance, Camille Berthelot, Matthieu Muffato, John E Collins, Sean Humphray, Karen Mclaren, Lucy Matthews, Stuart Mclaren, Ian Sealy, Mario Caccamo, Carol Churcher, Carol Scott, Jeffrey C Barrett, Romke Koch, Gerd-Jrg Rauch, Simon White, and Derek Stemple. Corrigendum: The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496, 04 2013.

[Hun81] Hunkapiller. R m hewick, Aug 1981.

[JAM97] PETER JAMES. Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly Reviews of Biophysics*, 30(4):279331, 1997.

[JM01] Lederberg Joshua and AT McCray. ome sweetomics-a genealogical treasury of words. *The Scientist*, 15(7):8–8, 2001.

[KNKA02] Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS / MS and Database Search. *Analytical Chemistry*, 74(20):5383–5392, 2002.

[LGT+18] Alexandre Lomsadze, Karl Gemayel, Shiyuyun Tang, Mark Borodovsky, Wallace H Coulter, Biomedical Engineering, Georgia Tech, Gene Probe, Computational Science, and Georgia Tech. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Research*, pages 1079–1089, 2018.

[MDA+15] Bart Mesuere, Griet Debyser, Maarten Aerts, Bart Devreese, Peter Vandamme, and Peter Dawyndt. The unipept metaproteomics analysis pipeline. *PROTEOMICS*, 15(8):1437–1442, 2015.

[NKKA03] Alexey Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, 75:4646–58, 10 2003.

[Pro] Software:proteinprophet. http://tools.proteomecenter.org/wiki/index.php?title=Software:ProteinProphet. Accessed 16-08-2019.

[RMS+11] Guus Roeselers, Erika K. Mittge, W. Zac Stephens, David M. Parichy, Colleen M. Cavanaugh, Karen Guillemin, and John F. Rawls. Evidence for a core gut microbiota in the zebrafish. *ISME Journal*, 2011.

[vdPD] Suzanne J. van der Plas-Duivesteijn. *Advancing Zebrafish Models in Proteomics*. PhD thesis, LUMC.

[vdPDMD+14] Suzanne J. van der Plas-Duivesteijn, Yassene Mohammed, Hans Dalebout, Annemarie Meijer, Anouk Botermans, Jordy L. Hoogendijk, Alex A. Henneman, Andr M. Deelder, Herman P. Spaink, and Magnus Palmblad. Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *Journal of Proteome Research*, 13(3):1537–1544, 2014. PMID: 24460240.