Utilizing Deep Learning to Uncover Antimicrobial Peptides and Investigating their Associations with Diverse Diseases

Samuel Chu
sjchu@ucsd.edu

Jonathan Dong
jqdong@ucsd.edu

Brianda Plascencia
bplascencia@ucsd.edu

Yiming Hao
y7hao@ucsd.edu

Rob Knight
rknight@ucsd.edu

Abstract

As pathogens start to become more resistant to our current antibiotics, we will soon need new antibiotics to fight various forms of bacteria. By 2050, drug-resistant pathogens are expected to be the leading cause of death in the world (O'Neill 2016). One way to create new antibiotics is to identify peptides, sequences of amino acids, that have antimicrobial properties. These are commonly known as antimicrobial peptides(AMP). These AMPs are known to regulate inflammation, kill certain types of cancer cells, and fight various infections and diseases. In this project, we use an Attention model(Vaswani et al. 2023) to classify peptides as AMPs or non-AMPs. We call the peptides that our model predicts to be AMPs: predicted AMPs. We then run this model on a dataset called FINRISK, which contains both DNA and health data on a randomly selected group of Finnish people. Using the output of our model and the health data in FINRISK, we found the correlation between having one of these predicted AMPs and having a disease recorded in FINRISK. In the 71 study participants we looked at, we found a moderately strong correlation between having 4 predicted AMPs and COPD, though these results are still preliminary in nature. We would need to run the model over all participants in FINRISK before being able to test our results in a wet lab.

Code: https://github.com/YimingHao0730/project2

1 Introduction

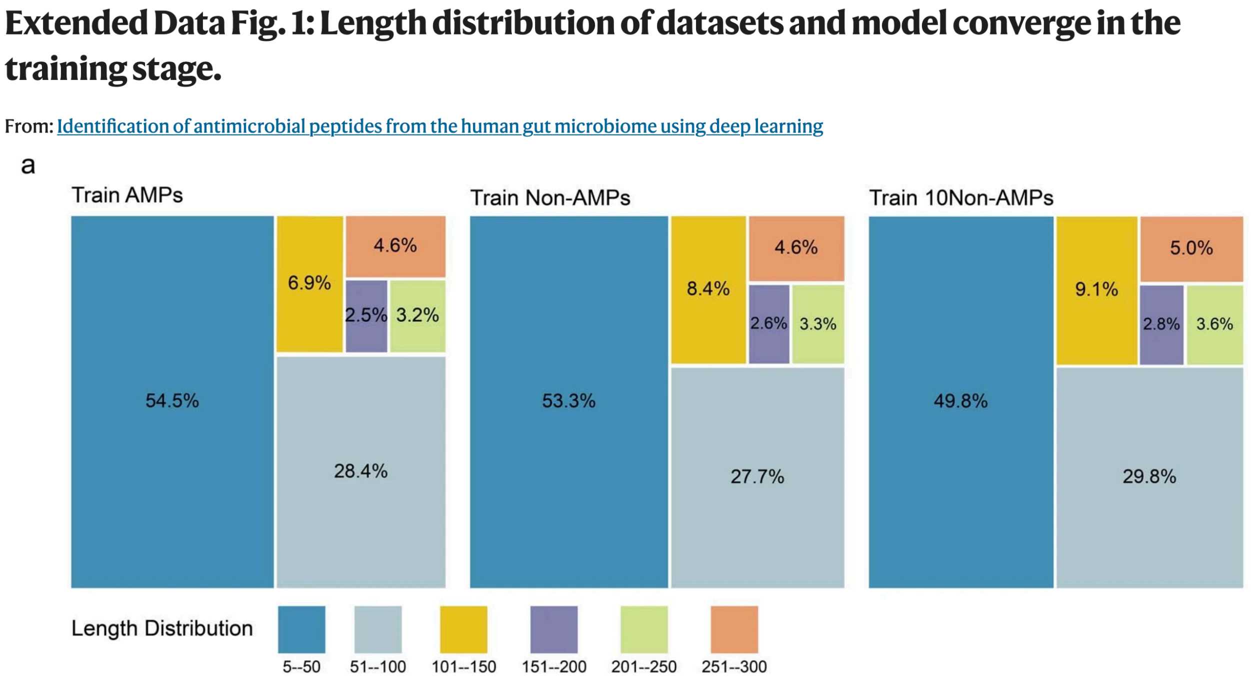
To match the rapid rise of drug resistant diseases, researchers have been trying to develop new antibiotics at a faster rate. Recently, the topic of antimicrobial peptides (AMPs) has been brought up as a possible solution to this problem because of their ability to reveal novel methods of killing harmful microbial pathogens. AMPs are oftentimes used to create new antibiotics, which are necessary to control the rise of drug-resistant patheogens. In our quarter 1 project, we replicated Ma et al. by using an Attention model to classify peptides as AMPs or Non-AMPs (Ma et al. 2022). Through our replication, we were able to obtain close to state of the art precision and recall values with some of our models that matched the results shown by Ma et al. Our Attention model had a precision of 70.36% and a recall of 91.24%. Using specifically the Attention model, we ran our model through data from the FINRISK dataset. We used the Attention model to predict which peptides in the FINRISK study participants were likely to be AMPs. Our goal in analyzing the FINRISK dataset is to understand if having any predicted AMPs has any correlation with having a disease. We have reason to believe that having certain AMPs may help reduce the risk of having some diseases because of their observed ability to kill or damage bacteria, viruses, fungi, and even cancerous cells (Reddy, Yedery and Aranha 2004). Finding an AMP that decreases the likelihood of developing a disease has the possibility to help researchers develop an antibiotic or cure to that disease.

2 Methods

This quarter we are using the AMP model that was developed in quarter 1 on a new data set called FINRISK. FINRISK was a study conducted in Finland that collected individual's DNA and statistics about their health such as which diseases they have and their weight. Given the DNA data of these individuals, our model will try to predict which AMPs a person in a study has. Our final analysis will aim to determine if the presence of any AMPs in a person are correlated with any health outcomes such as having heart disease or not having heart disease.

2.1 Model

The model takes in a list of peptides and outputs a corresponding probability value to each peptide, which represents the model's confidence that the peptide is an AMP. The Attention model we are using for this project was trained on a set of AMPs and non-AMPs. Ma et al. compiled a list of 10,322 known AMPs in late 2018, keeping 80% of them in the training set and 20% of them in the test set. The split of the non-AMPs between the train and test set seems to be a bit arbitrary at first, but our best explanation is that the current split is the result of the authors equalizing the distribution of lengths for the AMPs and non-AMPs in the training set. For example, in the training set, the percentage of the AMPs that are of length 5-50 is approximately the same as the percentage of non-AMPs that are of



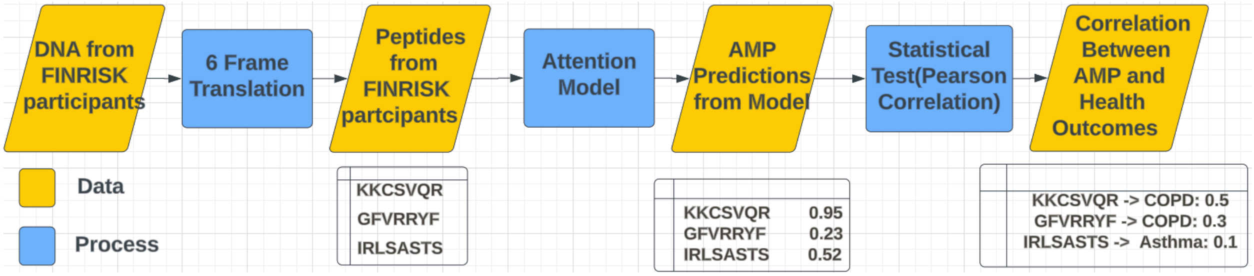
Extended Data Fig. 1: Length distribution of datasets and model converge in the training stage.

From: Identification of antimicrobial peptides from the human gut microbiome using deep learning

Figure 1: Distribution of Lengths in the Training Set

length 5-50(see Figure 1). The Attention model is trained on a dataset consisting of 8,290 AMPs and 74,838 non-AMPs, and the model is evaluated on a test set containing 2,032 AMPs and 2,908,751 non-AMPs. We used the trained weights of this Attention model in a file called att.h5 on Ma et al.'s Github. The Attention model is able to achieve a precision of 70.36% and a recall of 91.24% on amino acid sequences with length less than 50. The model consists of an embedding layer, a 1D Conv layer, 1D max pooling layer, and an Attention layer. The Attention model has become the state of the art deep learning method for understanding and generating text. In our case, however, our Attention model learns the meaning and relationship of amino acids instead of words.

2.2 Analysis on the FINRISK dataset



DNA from FINRISK participants
6 Frame Translation
Peptides from FINRISK partcipants
Attention Model
AMP Predictions from Model
Statistical Test(Pearson Correlation)
Correlation Between AMP and Health Outcomes
Data
Process

KKCSVQR
GFVRRYF
IRLSASTS

KKCSVQR	0.95
GFVRRYF	0.23
IRLSASTS	0.52

KKCSVQR -> COPD: 0.5
GFVRRYF -> COPD: 0.3
IRLSASTS -> Asthma: 0.1

Figure 2: Project Workflow

2.2.1 FINRISK dataset

The FINRISK dataset was a study conducted the by Finnish government that contains DNA and health information of randomly selected Finnish people. Since the FINRISK dataset deals with confidential human data, we were only allowed to access the data from Barnacle2, the Knight Lab's supercomputer cluster.

2.2.2 Data Preprocessing

Since our model requires peptides, amino acid sequences, and the FINRISK data only contains DNA sequences of study participants, we have to do a translation from DNA sequences to amino acid sequences. To accomplish this, we used EMBOSS to do six frame translation from DNA sequences to amino acid sequences. Six frame translation reads the DNA across three potential overlapping frames forward, as well as the three frames on the reverse complement of the nucleotide sequence. This allows us to identify all possible codons from a nucleotide sequence. We then look at the Open Reading Frames (ORFs) of these amino acid sequences to find the valid sequences that lie between a start and a stop codon. After six frame translation and ORF extraction, we have a list of peptides from each study participant that is ready to be fed into our Attention model.

2.3 Prediction

Once our data is preprocessed, we pass the data into our Attention model, which predicts how likelihood of each peptide being an AMP. If the probability that a peptide is an AMP is over our threshold value, then we will predict that it is an AMP. Since, our input file to our model can be upwards of 15 million peptides long, our model would oftentimes predict 4 million of them to be AMPs, when a threshold of 0.5 is used. Therefore, to drastically reduce the number of AMPs we get per study participant, we used a threshold of 0.99999.

2.4 Analysis of Each AMP's Correlation to A Health Condition

Now, for every study participant in FINRISK, we have a list of their peptides that we expect to be AMPs. We also have their health data from FINRISK, which tells us which diseases a study participant has or does not have. With this data, we perform a Pearson Correlation Coefficient test between every AMP and some diseases recorded in FINRISK.

3 Results

The Pearson Correlation Coefficient test describes the strength and direction of a linear relationship with the correlation coefficient, r. R ranges from -1 to 1.

Table 1: Number of Peptides Per Category

Disease	Moderately Correlated (r>=0.5)	Lightly Correlated (0.5>r>0.3)	Negatively Correlated (0.0>r)
Type 2 Diabetes	0	3	0
COPD	4	0	0
Asthma	0	3	0
HDL	0	3	3

Peptides Moderately Correlated With COPD (r>=0.5):

1. GFVRRYFGYKSGILCRRGCVCRGWKRK
2. IRLSASTSICKVSCTVSDKACCCSGGSLSNTGVCCSCKNSGTC
3. CPAFSGHCHTPWGVCRPAMCRCAE
4. KKCSVQRCTFSYAKKDGKCKGMFRVE

Peptides Lightly Correlated With Type 2 Diabetes (0.5>r>0.3):

1. PYRKWCNNSCCVEGVAVWCPNCDNG
2. AERIPRCDQQAAGQGCGRGVCRFRRCGGKAW
3. WKWPGNSIRCSAVRRQTWYRSAWCCSAWTW

Peptides Lightly Correlated With Asthma (0.5>r>0.3):

1. LWAVCRKVCRR
2. CYRNRLCCSSCSKG
3. IYGSFKRRFGCCHLRNTC

Peptides Lightly Correlated With HDL (0.5>r>0.3):

1. CRIFKCIISICRK
2. NYFRKGFCPRNECAVH
3. YPIRGTCIKTFC

Peptides Negatively Correlated With HDL (0.0>r):

1. RDLYRSNICCIRHGYC
2. WRQGLCRWGGCR
3. RPLQHRLQRFGKKIRRRNSCQVPS

4 Discussion

The strongest correlation between a peptide and a disease occurs between the 4 peptides listed above and COPD. The Pearson Correlation Coefficient of the 4 peptides and COPD vary from 0.5308 to 0.5604. These results need to be investigated further but may suggest that having one of these 4 peptides could increase one's chance of having COPD. We want to recreate these peptide sequences in a wet lab to hopefully discover new AMPs and to better understand COPD. However, as mentioned before, we will need to run all the participants in FINRISK through our model and through our statistical analysis before moving forward to wet lab testing.

5 Future Goals

If we have more time to work on this project, we would like to be able to run all the participants in FINRISK through our model. Right now we are only the peptides of 71 study participants. We would also like to do a Pearson correlation test between every predicted AMP and every disease in FINRISK. Right now, we are only looking at a handful of diseases. If we find significant correlation between a predicted AMP and a disease after running our model through all of FINRISK, we will try to work with a wet lab to sequence the peptide. The peptides that have some correlation with COPD are not known, so it could be useful to get them sequenced. Additionally in the future, it might be useful to train a new model on new AMP data since the model we are currently using was trained on data from 2018. Lastly, we might want to run our model on other datasets similar to FINRISK to see if we can make any discoveries there. The FINRISK dataset is specific to Finnish people, so we might see new or differing results when looking at people from other countries or backgrounds.

References

Ma, Yue, Zhengyan Guo, Binbin Xia, Yuwei Zhang, Xiaolin Liu, Ying Yu, Na Tang, Xiaomei Tong, Min Wang, Xin Ye et al. 2022. "Identification of antimicrobial peptides from the human gut microbiome using deep learning." *Nature Biotechnology* 40 (6): 921–931

O'Neill, Jim. 2016. "Tackling drug-resistant infections globally: final report and recommendations.": 2740–2747

Reddy, K.V.R., R.D. Yedery, and C. Aranha. 2004. "Antimicrobial peptides: premises and promises." *International Journal of Antimicrobial Agents* 24 (6): 536–547. [Link]

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. "Attention Is All You Need."