Classification, Prediction and Analysis of Type VI Secreted Effector Proteins

{ Rishika Sen, Losiana Nayak & Rajat K. De }
Machine Intelligence Unit, Indian Statistical Institute, Kolkata

1. What are Secretion Systems?

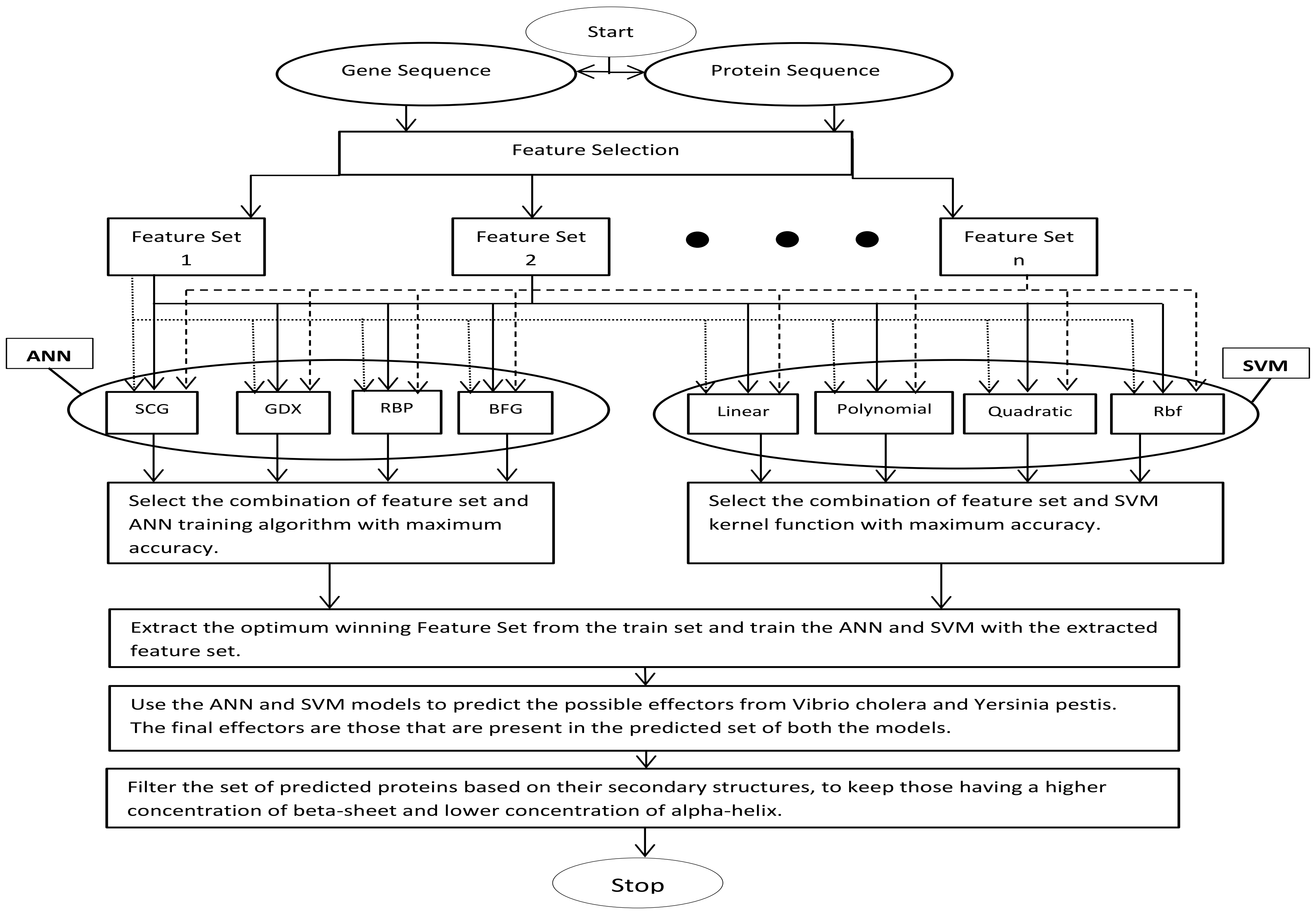
1. Pathogens have numerous ways of transporting protein between locations via secretion apparatus [1].
2. Bacterial secretion apparatuses can be divided into different types, based on their structures, functions, and specificity.
3. Secretion systems are used by bacterial pathogens to manipulate the host and establish a replicative niche [2], often ending in infection.

2. What is the need of classification?

1. Effector proteins of bacteria infect their hosts by secretion systems (SS).
2. Effector proteins of the T6SS of many species are yet to be discovered.
3. No signature pattern have yet been discovered to differentiate T6 effector proteins from the non-effectors ones.
4. Such signature patterns will speed the discovery of T6 effector proteins in many gram-negative bacteria.

3. Methodology

Features of nucleotide sequences are 4 mono-nucleotide, 16 dinucleotide and 64 trinucleotide frequencies. The protein feature matrix has been made up of 20 single amino acid frequencies, 400 dipeptide frequencies and 11 physico-chemical property based features. We have applied Sequential Forward Selection (SFS) strategy on the feature set of the experimentally verified T6 effector proteins to rank our features according to their significance. According to these ranks, we have created sets of 8, 16, 24, 32 and 40 most significant features. We have compared the performance of the different classifiers with each of these sets and the set with maximum accuracy and stability has been used for prediction of effector proteins in *Vibrio cholerae* and *Yersinia pestis*.



Start
Gene Sequence
Protein Sequence
Feature Selection
Feature Set 1
Feature Set 2
Feature Set n
ANN
SCG
GDX
RBP
BFG
SVM
Linear
Polynomial
Quadratic
Rbf
Select the combination of feature set and ANN training algorithm with maximum accuracy.
Select the combination of feature set and SVM kernel function with maximum accuracy.
Extract the optimum winning Feature Set from the train set and train the ANN and SVM with the extracted feature set.
Use the ANN and SVM models to predict the possible effectors from Vibrio cholera and Yersinia pestis. The final effectors are those that are present in the predicted set of both the models.
Filter the set of predicted proteins based on their secondary structures, to keep those having a higher concentration of beta-sheet and lower concentration of alpha-helix.
Stop

Figure 1: Methodology pipeline depicting the flow of the procedure

4. Classifier Result

The performance of the training algorithms is as follows:

Class	**SCG**	**RP**	**BFG**	**GDX**
ANN	90.02	90.74	88.53	80.21
Class	**Linear**	**Polynomial**	**Quadratic**	**RBF**
SVM	93.3	95.40	93.5	91.6

Table 1: Summary of classifier performance

The ANN trained with SCG and SVM with linear kernel have predicted T6 effector proteins with maximum stability. Out of **2267** proteins of *Vibrio cholerae*, **611** proteins have been predicted by both ANN and SVM to be putative T6 effector proteins. For *Yersinia pestis*, out of **2736** proteins, **1058** proteins have been predicted to be possible T6 effector proteins.

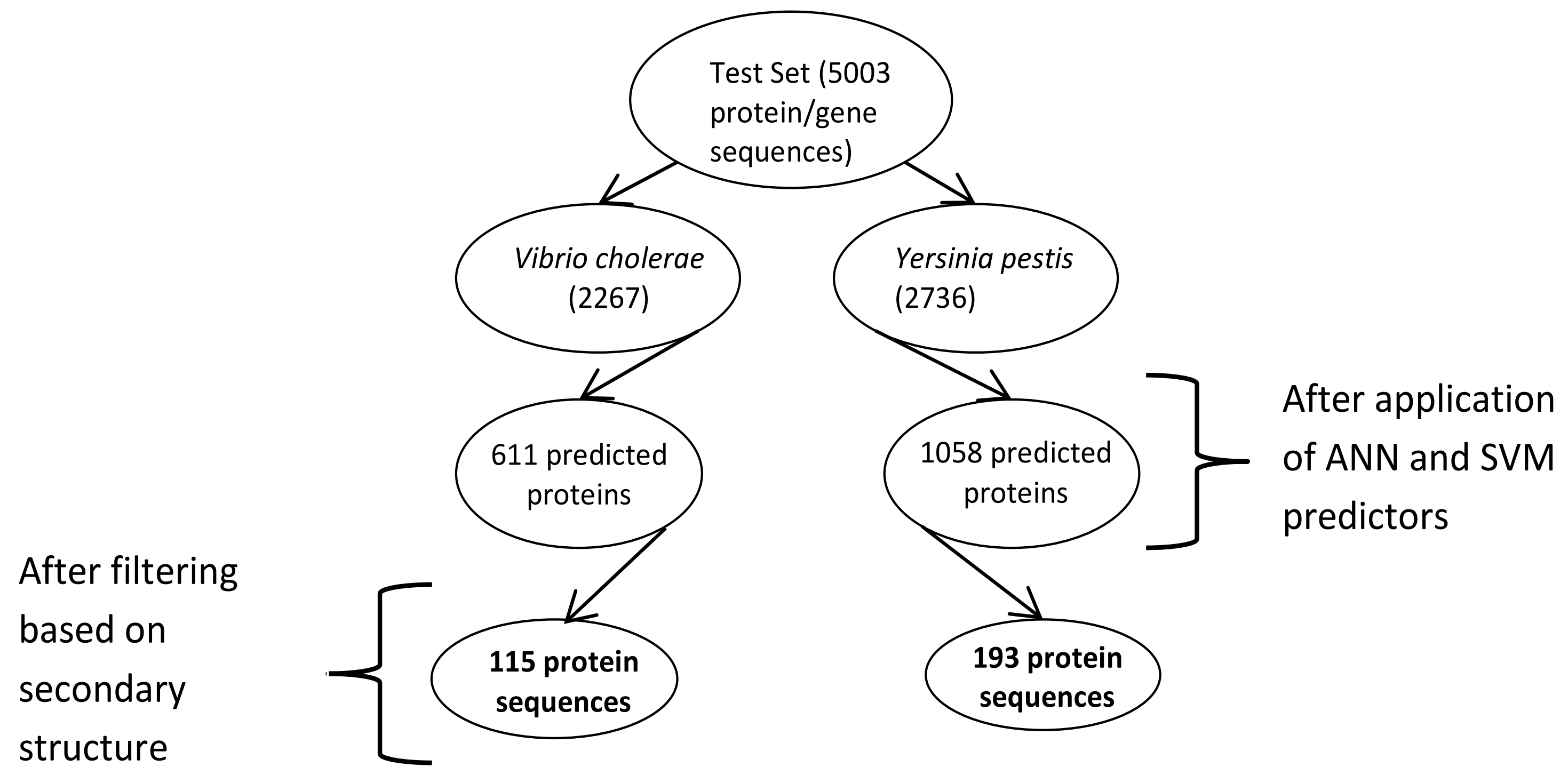
5. Secondary structure based filtering

Table 2 highlights secondary structure based findings of the predicted effector proteins. The ML based predicted effector proteins have been filtered based on secondary structure of proteins of *Vibrio cholerae*, **115** proteins have been found with higher concentration of sheets than helices. Likewise, **193** proteins of *Yersinia pestis* have been found with higher concentration of sheets than helices, as was observed in the experimentally verified T6 effector proteins. These **115** and **193** proteins have a higher chance of being T6 effector protein, since they have a high structural similarity with the experimentally verified effector protein list.

Class	**Coil (%)**	**Helix (%)**	**Sheet (%)**
Effector	47.59	19.22	33.17
Non-Effector	60.16	29.55	10.28

Table 2: Composition of secondary structure components

6. Summary



Test Set (5003 protein/gene sequences)
Vibrio cholerae (2267)
Yersinia pestis (2736)
611 predicted proteins
1058 predicted proteins
After application of ANN and SVM predictors
After filtering based on secondary structure
115 protein sequences
193 protein sequences

Figure 2: Detailed numerical and species-wise information of the filtering pipeline leading to a more potent set of effector proteins.

We have predicted a set of 611 T6 effector proteins out of 2736 proteins from *Yersinia pestis*. We have also predicted a set of **1058** T6 effector proteins out of **2267** proteins of *Vibrio cholerae*. Among them, **193** proteins of *Yersinia pestis* and **115** proteins of *Vibrio cholerae* have been found with higher percentage of sheets and lower percentage of helices than the non-effector proteins.

These findings have strengthened our prediction, as higher percentage of sheets and lower percentage of helices is a feature often found in other types of effector proteins. The classifier designed here can be further extended to predict effector proteins in other gram-negative (pathogenic) bacteria.

Availability of more experimentally validated T6 effector proteins will generate a better understanding of their structure and functionality, which in turn will strengthen our prediction pipeline.

7. References

[1] Sen et.al. A review on host–pathogen interactions: classification and prediction. *European Journal of Clinical Microbiology & Infectious Diseases*, 35(10):1581–1599, 2016.

[2] Costa et. al. Secretion systems in gram-negative bacteria: structural and mechanistic insights. *Nature Reviews Microbiology*, 13(6):343–359, 2015.

Thank you!

Contact: rishika21sen@gmail.com

Acknowledgment: Losiana Nayak acknowledges University Grants Commission, India for providing her a UGC Post-Doctoral Fellowship (No. F.15-1/2013-14/PDFWM-2013-14-GE-ORI-19068(SA-II)).