In *Silico* Approach for Prediction of Antifungal Peptides

Year Published : 2018
Course name and code : BDMH - BIO543

Presented By :

Meghal Dani (MT17144)
Shubhi Tiwari (MT17057)

Introduction - About the Paper

The paper describes **in silico models** developed **using** wide range of **peptide features** to **predict antifungal peptides (AFPs).**

The paper employs **Machine Learning Techniques** to derive rules from experimentally validated **AFP** and **non-AFPs to discriminate these two classes of peptides.**

The Need

Why work on this problem?

- High morbidity and mortality rate observed due to invasive fungal infections.

- AMPs can be used to treat fungal infections but lack of specificity reduces their potential.

Our Approach to the Project

The process followed

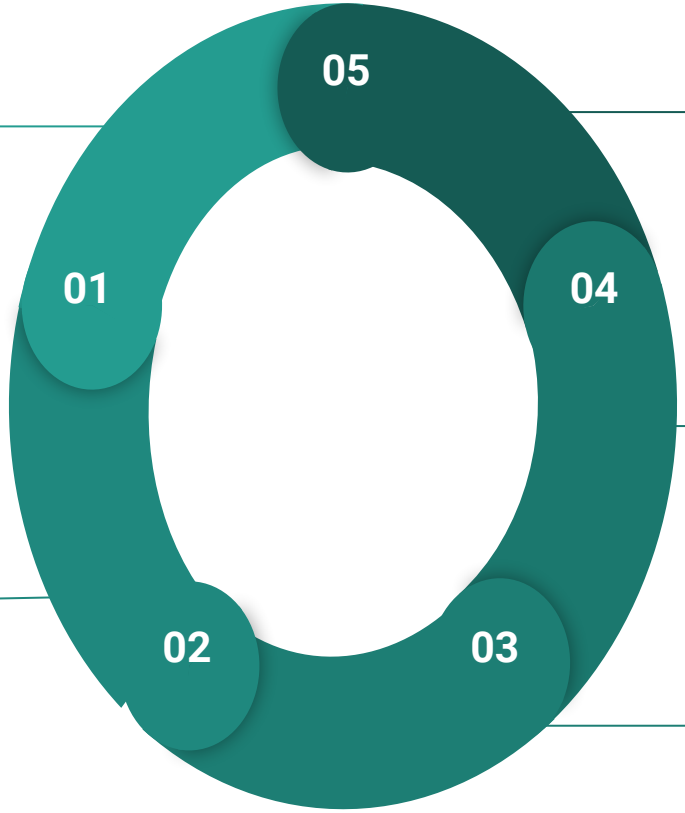
Workflow

Dataset Retrieval

- Antifp_DS1
- Antifp_DS2
- Antifp_Main

Input Features

- Amino Acid Composition
- Dipeptide Composition
- Split Composition
- Binary Profiles
- Amino acid comp with mass, charge and PI



Performance measure

Performance of each model was measured and reported on Validation dataset

Model Training

- **By Authors** - SVM, Random Forest, J48, Naive Bayes
- **Extra** - ELM, Adaboost, KNN, Ensemble on Random Forest.
- **Disregarded** : Neural Networks, Rotation Forest, Ensemble on SVM

Internal and External Validation

- Internal : 5 Fold Cross Validation
- External : External Validation data used

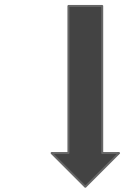
Step 1 : DataSet

- Dataset downloaded from http://webs.iiitd.edu.in/raghava/antifp/algo.php
- Consists of **1459** unique sequences of AFPs and non-AFPs of various lengths.
- Dataset was already divided into **80-20 %** split
  - **Training dataset** - 1168 positive and negative sequences
  - **Validation Dataset** - 291 positive and negative sequences.
- Data was available in .txt file format

Step 2 : Converting txt files to FASTA

- **Input** : txt files of peptide sequences

```
HDEF
KDEL
RWRW
AHKCIC
FPAHKC
```



Python Scripts

- **Output** : FASTA files

```
>1
HDEF
>2
KDEL
>3
RWRW
>4
AHKCIC
>5
FPAHKC
```

Step 3 : Extracting features from peptide sequences

- Input : FASTA files of peptide sequences
- Output : csv files of generated features

Features Generated

- Amino Acid composition
- Dipeptide Composition
- Split Composition
- Binary Profile
- Physicochemical Properties - Mass, Charge and PI values

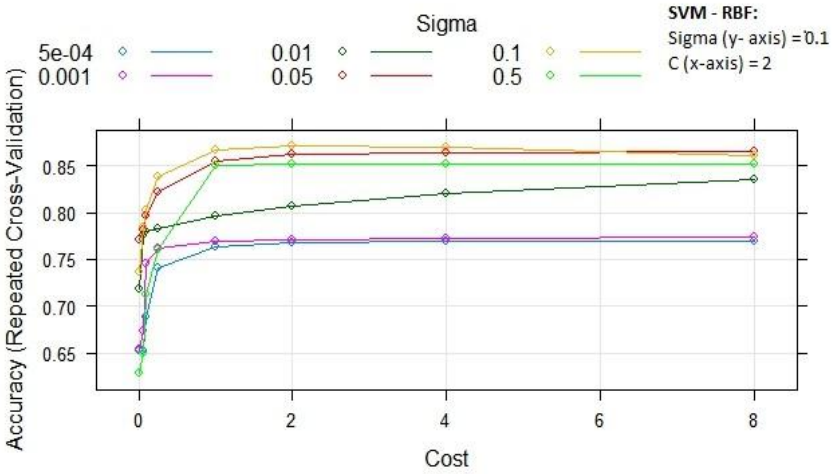
Step 4 : Trying out various classification Techniques

- Input : feature wise csv files
- Tried techniques applied in the paper
- Tried a few new techniques
- Used R
- Applied 5 fold cross validation

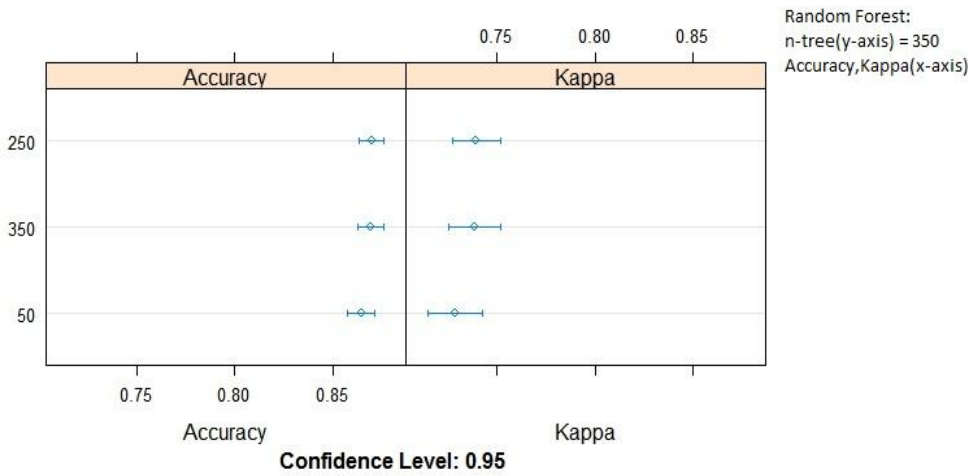
	Techniques	Applied in paper	Applied by Us	
			Selected	Rejected
01	SVM Radial	Yes	Yes	-
02	Random Forest	Yes	Yes	-
03	J48	Yes	Yes	-
04	Naive Bayes	Yes	Yes	-
05	ELM	-	-	Yes
06	Adaboost.M1	-	Yes	-
07	KNN	-	Yes	-
08	Neural Network	-	-	Yes
09	Rotation Forest	-	-	Yes
10	Ensemble (SVM radial, RF, J48, Adaboost.M1)	-	Yes	-

Step 5 : Model parameter tuning

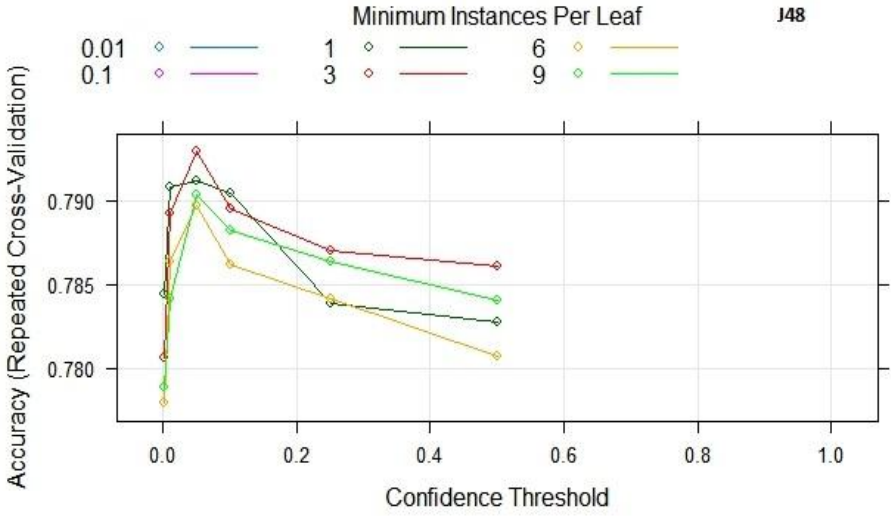
- Models' parameters were tuned to determine the best possible parameters for each model



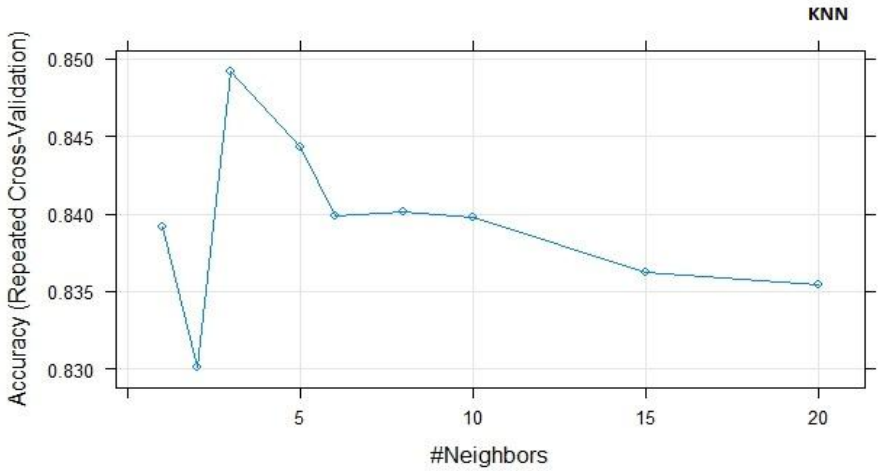
SVM -RBF



Random forest



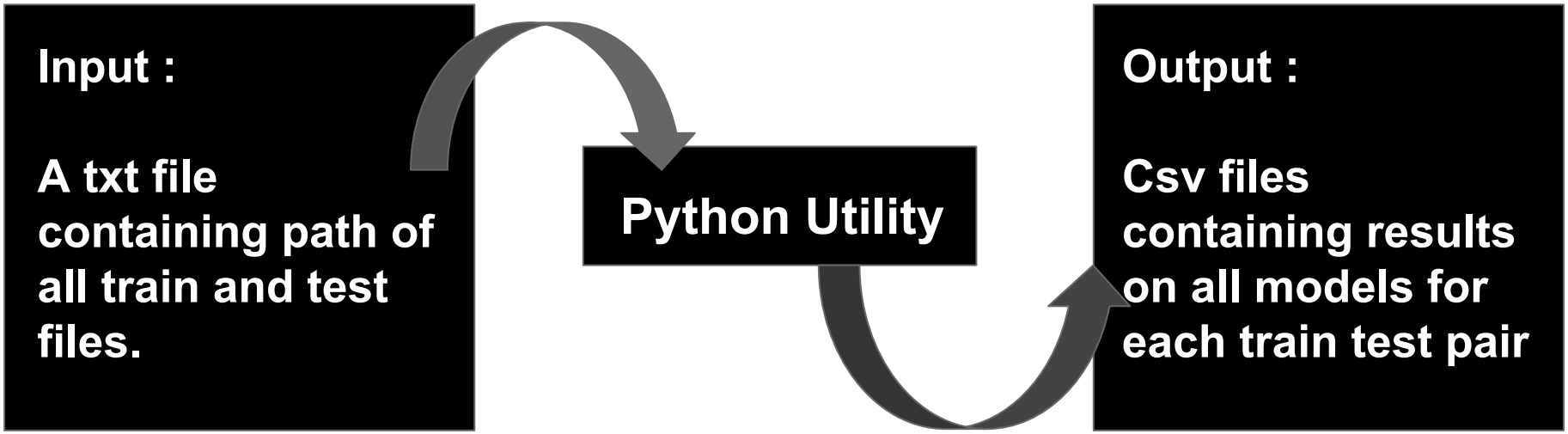
J48



K- Nearest Neighbours

Step 6 : Training and Testing the models

- For each of the generated feature files
  - Models were trained on the 80% train data
  - Models were tested on the 20% validation data
- Created a python utility to test all models for a set of input data in one go.



Results

Dataset : Antifp_DS2
Feature : amino acid composition

Model	Train	As reported in paper				As obtained by us			
		Acc	Sen	Spe	ROC	Acc	Sen	Spe	ROC
Naive Bayes	Train	85.83	84.42	87.24	0.91	88.27	93.66	82.8	0.88
	Test	84.36	83.16	85.57	0.90	87.45	91.06	83.84	0.91
RF	Train	91.65	91.95	91.35	0.97	100	100	100	1
	Test	87.29	87.97	86.60	0.93	91.5	92.78	90.3	0.91
SVM Radial	Train	92.81	93.24	92.38	0.97	98.0	98.45	97.6	0.98
	Test	90.38	90.72	90.03	0.96	90.7	91.4	90.03	0.90

Model	Train	As reported in paper				As obtained by us			
		Acc	Sen	Spe	ROC	Acc	Sen	Spe	ROC
J48	Train	87.97	88.87	87.07	0.91	92.46	91.5	93.40	0.92
	Test	86.77	87.97	85.57	0.90	86.25	82.81	89.69	86.25
Adaboost.M1	Train	-	-	-	-	100	100	100	1
	Test	-	-	-	-	88.83	89.00	88.65	0.88
KNN	Train	-	-	-	-	91.69	91.26	92.12	0.91
	Test	-	-	-	-	89.86	87.97	91.75	0.89

Results - Ensemble

Models : SVM Radial, RF, J48, AdaBoost.M1
Stacking : RF

Dataset : Antifp_DS2
Features : Amino Acid Composition

Accuracy - Training

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max. NA's
svmRadial	0.89	0.90	0.91	0.91	0.92	0.93
rf	0.90	0.91	0.92	0.92	0.92	0.95
J48	0.85	0.86	0.87	0.87	0.88	0.91
AdaBoost.M1	0.89	0.90	0.91	0.91	0.92	0.95

Accuracy - Validation

SVM Radial	**RF**	**J48**	**AdaBoost. M1**	**Ensemble**
96.22	96.78	87.89	95.32	**96.58**

Conclusion

- In this project, we developed ML models to distinguish AFP and non-AFP sequences.
- We generated features from peptide sequences on which the models were trained.
- We were able to generate results similar to that of the original paper
- We created an ensemble of the four best performing models (SVMRadial , J48,RF, AdaBoost.M1), combined using random forest which out performed all previous models.

Thanks