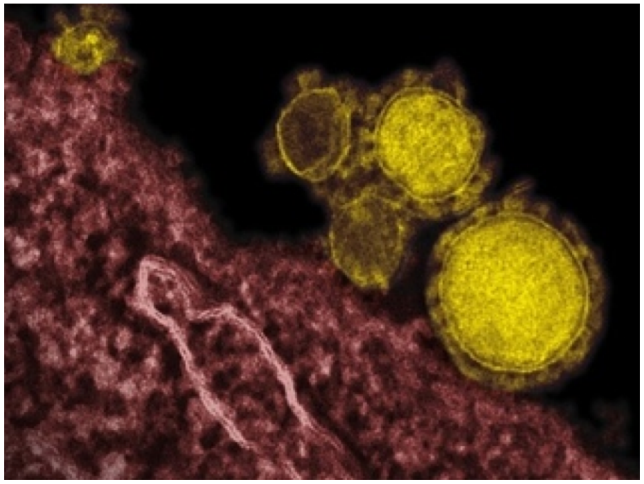
Prediction of host-pathogen protein-protein interactions using machine-learning

Cristian Loaiza
Plant Science MS Student
Advisor: Dr. Rakesh Kaundal

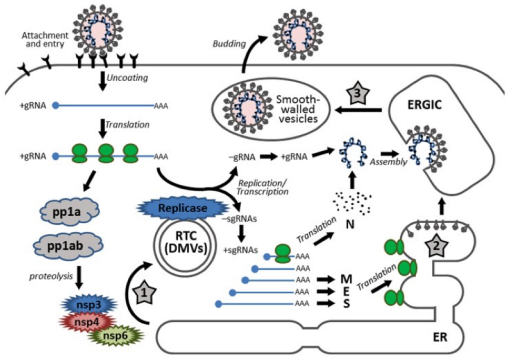
Introduction

Host-Pathogen interactions (HPI)

- In biology, a host is an organism in which a pathogen obtain its nutrition and/or shelter.
- A pathogen is an organism that cause disease (Viruses, Bacterias, Protozoans, Fungi, and Parasites).
- Examples of host-pathogen interactions:
  - Human/HIV-1 (AIDS)
  - Human/Influenza virus
  - Wheat/*Puccinia spp.* (Rust)



credit: NIAID



credit: (Fung and Liu, 2014)

Why is important to study HPI?

- Each year **millions of people die** due to infectious diseases (WHO 2017).
- About 65% of U.S. crop losses are due to non-indigenous pathogens, amounting to an estimated cost of >**$40 billion a year**.
- Infectious diseases are being discovered at a higher rate than at any time in history (WHO 2007).
- **Host-pathogen protein-protein interactions** (HPI) play a crucial role in infectious processes among with the multiple environmental factors present in nature (Casadevall, 1999).

HPI have been studied to find potential genomic targets for the development of **novel drugs, vaccines and other therapeutics** (Briken, 2008).

Why computational prediction?

- To identify host-pathogen interactions experimental proteomics analysis are used:
  - Yeast two hybrid (Y2H)
  - Affinity-purification mass spectrometry (AP-MS)
  - Immuno-coprecipitation (Co-IP)
- Those methods are **time-consuming** and **expensive** if are performed at a large scale.
- **Computational prediction solves the problem of both time and cost** from the classical proteomic methods, either by the use of databases templates, machine learning models or sequences similarities.

Why machine-learning?

Is ideal for complex systems	• Machine learning have been proved efficient to summarize complex systems.
Now, we have the resources	• Recently, efforts have been made to collect most of the data present in the literature about HPI. (Ammari, 2016)
Novel approach for HPI	• Not a comprehensive study has been done yet exploring machine learning models for HPI prediction.

Question & Goal

Question: Are machine learning methods suitable to predict host-pathogen protein-protein interactions?

Goal: Assess the effectiveness of machine learning models to summarize host-pathogen protein-protein interactions.

Methodology

Support-vector machines (SVM)

- SVM was the algorithm selected for this classification problem.
- SVM is a supervised machine learning algorithm.
- Keys to understand supervised machine learning algorithms:
  - Learn from given training examples.
  - Each training example is marked belonging to one or the other of two categories. (e.g: **positive** and **negative**)
  - The model learn how to classify new data points into one of those two categories.

Ideal for HPI prediction problem, because we want to classify a protein pair into two categories (**interacting** or **not interacting**).

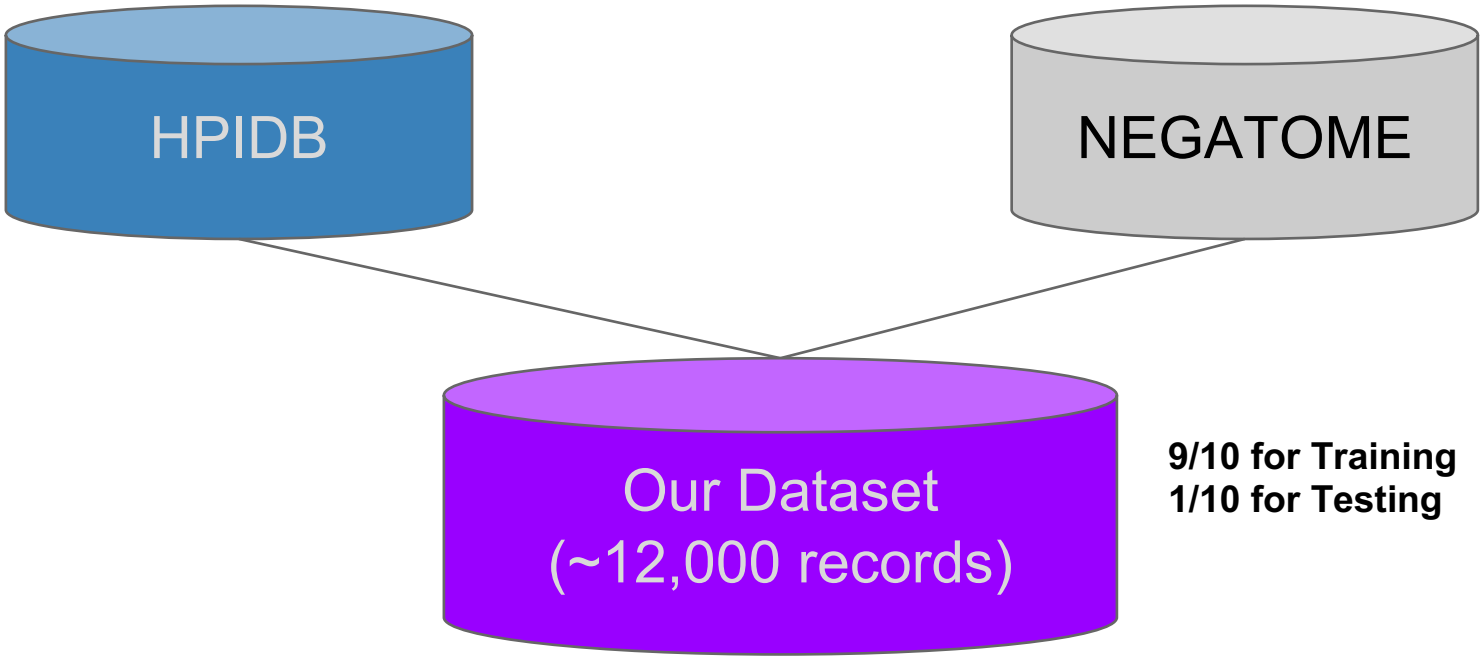
Data gathering

Positive samples

(protein-protein **interactions**)

Negative samples

(protein-protein **non-interactions**)



Protein-protein representation

- In computational biology, a protein is a sequence of amino acid letters.

- A protein-protein interaction is composed of two sequences of amino acid letters.
- In order to build an SVM model, we need to represent a protein-protein interaction with a vector of a fixed size.
- The vector needs to conserve at most the protein pattern and the amino acid sequence order (Minimum information loss).

Protein features

- Four protein features were selected for this analysis:
  - Geary Autocorrelation composition (480 features per record).
  - Dipep composition (800 features per record).
  - Conjoint Triad composition (686 features per record).
  - Quasi-order composition (100 features per record).

- Comparing the performance among different features could guide us to a better understanding of what patterns are important to emphasize, in order to characterize properly those protein-protein interactions datasets.

Analysis Pipeline

- Transform training/ testing dataset into the four features representations. (protr R package)
- Build an SVM model for each training set (e1071 R package) and tune SVM parameters (cost, sigma and kernel).
- Compare the models.
    - K-fold cross validation (Training dataset).
    - Validation testing (Testing dataset).

Results

K-fold cross validation

	AC (Geary)	Conjoint Triad	Dipep	Quasi-order
Sensitivity	0.954	0.898	0.897	**0.996**
Specificity	0.982	**0.995**	0.980	0.875
Accuracy	**0.978**	0.946	0.947	0.969

Training

- Quasi-order model classified better the positive samples.
- Conjoint Triad model classified better the negative samples.
- Geary model performed best.
- Overfitting?

Validation testing

	AC (Geary)	Conjoint Triad	Dipep	Quasi-order
Sensitivity	0.967	0.922	0.928	**0.995**
Specificity	0.88	0.972	**0.980**	0.561
Accuracy	0.926	0.947	**0.954**	0.778

Testing

- Quasi-order model classified better the positive samples.
- Dipep model classified better the negative samples.
- Dipep model performed best.

Are machine learning methods suitable to predict host-pathogen protein-protein interactions?

- SVM has been proved **useful** to predict host-pathogen protein-protein interaction according to our analysis (Best model achieved 95% of accuracy).
- Machine-learning could be used to identify novel effectors in infectious diseases in unknown host-pathogen systems.
- However, there is not clarity in which of the features is doing a better job representing the protein-protein interactions attributes.

Conclusions & Future Work

Conclusions

- We have assess the effectiveness of **support-vector machines** models to summarize host-pathogen protein-protein interactions.
- Most of the features seem to be doing a decent job to characterize the variability between the positive database (HPIDB) and the negative (Negatome).
- We hope that from our comparison, the fundamentals in the prediction of host-pathogen interaction using machine learning techniques will be settled.

Future work

- Perform a comparison among different machine learning methods: SVM, Artificial Neural Networks and Convolutional Neural Networks.
- Generate a more suitable non-interaction dataset, Negatome is an intra-species database.

Acknowledgments

STAT 6310 Final Project Members:
Jonathan Medri
Chatumaverdi Ediweera

Kaundal Lab members:
Matthew Lister (Linux Researcher)
Naveen Duhan (PhD Student)

Funding support:

- **USDA CDRE grant**
- **USU (PSC, CIB) - Kaundal Bioinformatics Laboratory**

Thank you for your kind attention!

Auxiliar Slides

Machine-learning

credit: https://xkcd.com/1838/

Build and Tuning of models

Model parameters

	AC (Geary)	Conjoint Triad	Dipep	Quasi-order
Kernel	Radial	Radial	Radial	Radial
Cost	5	5	1.5	5
Sigma	0.01	0.01	0.01	0.01

- Dipep was the model that took more time to train.
- Quasi-order was the model that took less time to train.
- Radial kernel.

Metrics to compare the models

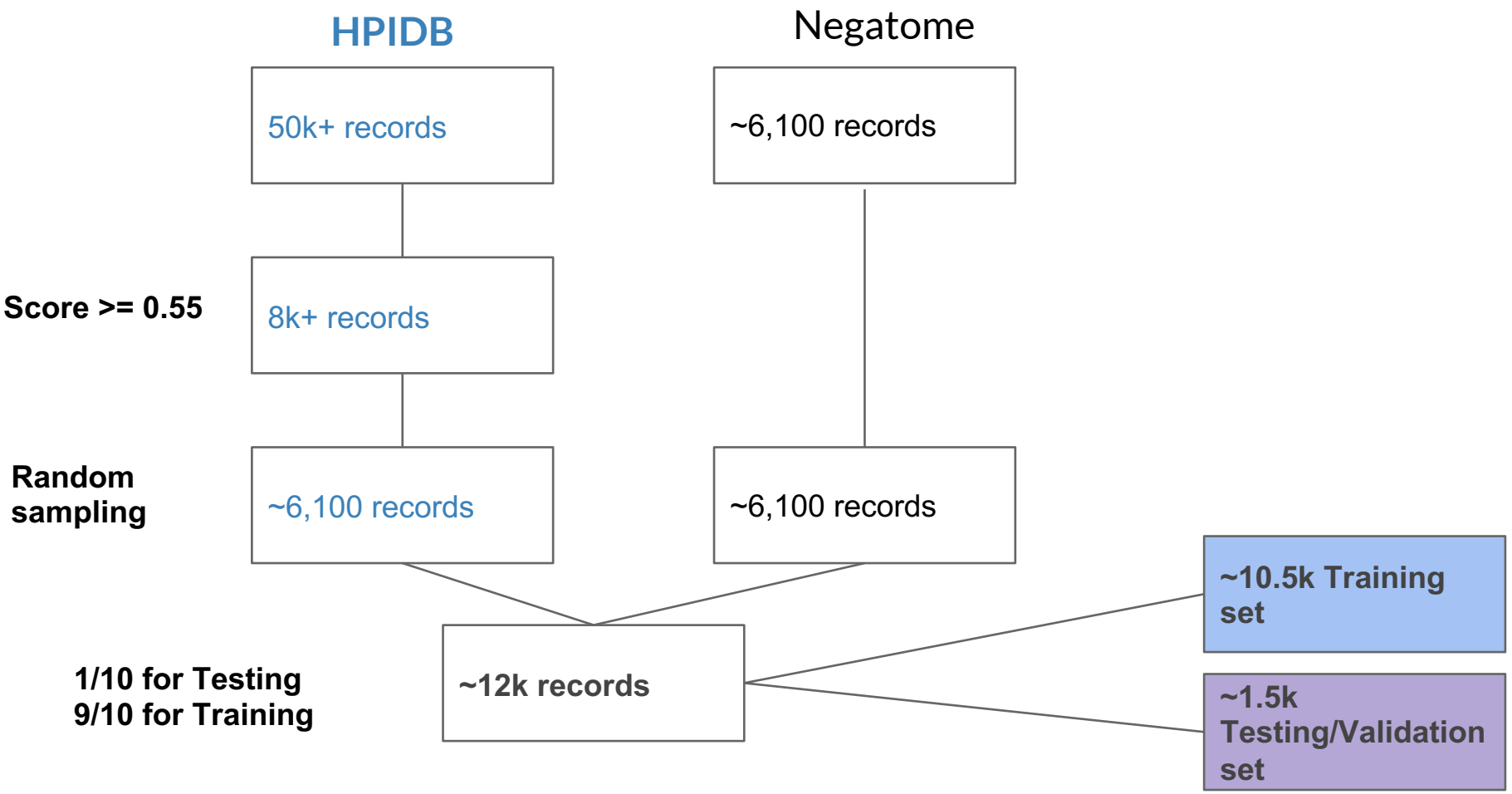
- True Positives (TP): Positive samples predicted as positive.
- False Positives (FP): Negative samples predicted as positive.
- True Negatives (TN): Negative samples predicted as negative.
- False Negatives (FN): : Positive samples predicted as negative.

Sensitivity = TP / (TP + FN)

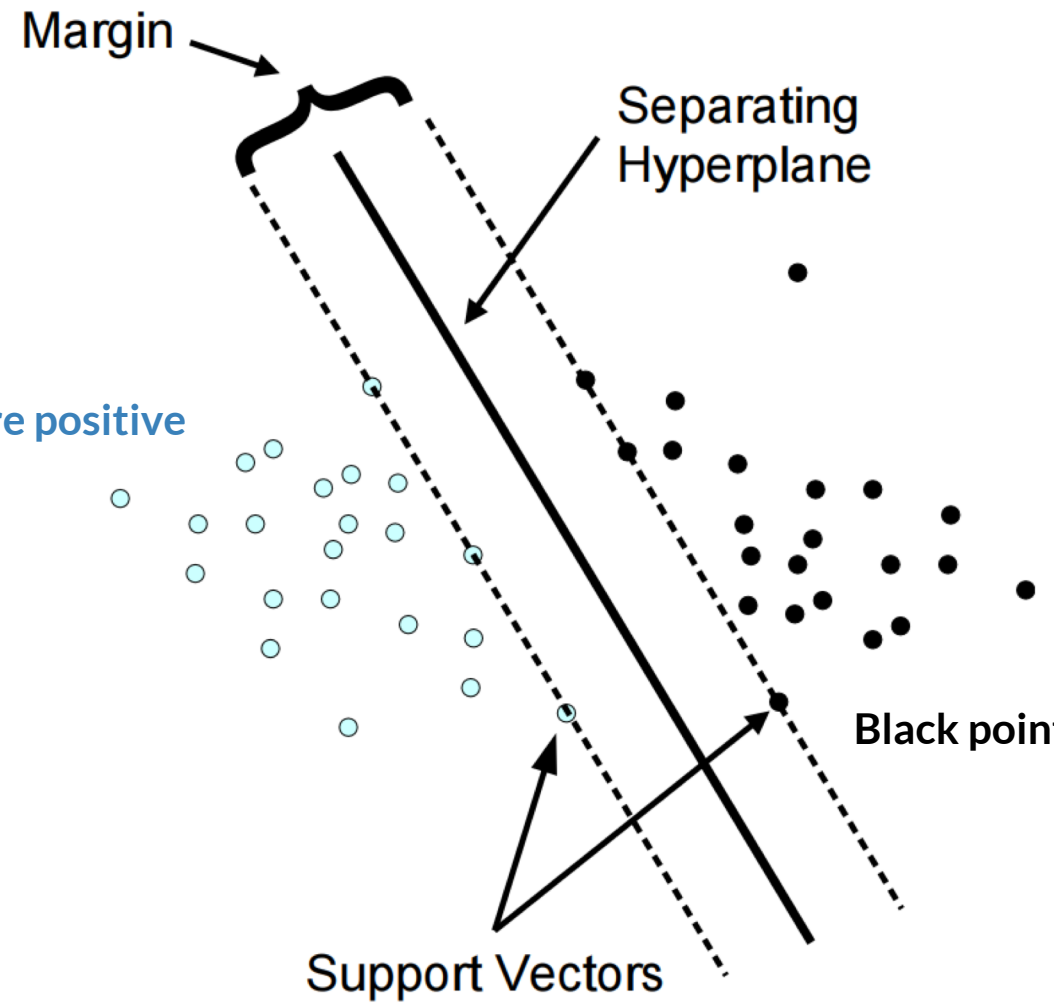
Specificity = TN / (TN +FP)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Training/ Testing Datasets



How SVM works?



Blue points are positive

Black points are negative

credit: vignette for the R package e1071