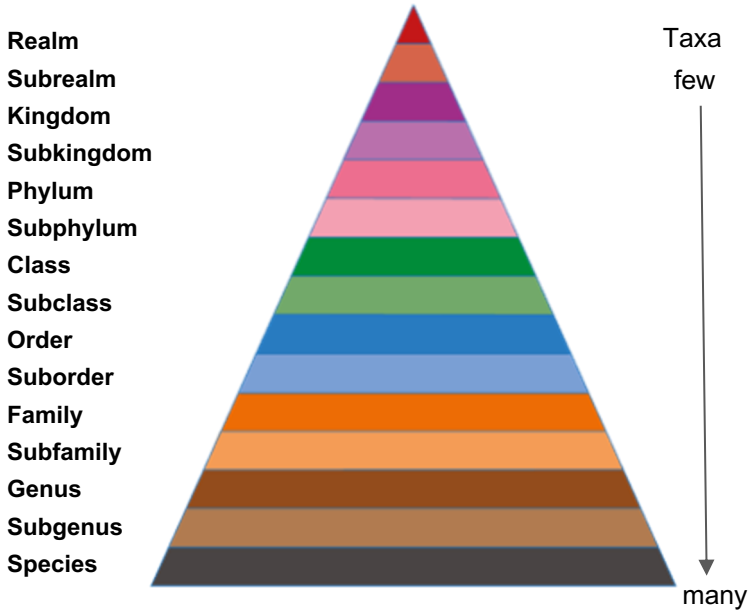
Classification of Viral Pathogens Based on Multiple Genomic Signatures

Anoushka Bhat[1], Esha Ananth[2], Rishov Chatterjee[3], Srisairam Achuthan[3] PhD, Samir Courdy[3]

[1]Diamond Bar High School, Diamond Bar, CA
[2]Portola High School, Irvine, CA
[3]Research Informatics Division, Center for Informatics, City of Hope, CA

Introduction

- Taxonomic classification : finding the identity of a certain virus[1]
- For unknown, potentially harmful pathogens, classification can help uncover patterns from closest known pathogens
- 10 taxonomic levels for viral genome, each has 1 or more sublevels



Adapted From: https://talk.ictvonline.org/

[1] Richards R, Biological Classification: A Philosophical Introduction, Cambridge University Press, 2016

Objective: Sars-Cov2 Sequence Classification

- Simplify classification and prevent data leakage by creating a new feature to classify Sars-Cov-2 sequences into a sublevel at each of the 9 out of 10 taxonomic levels[3].

Taxonomic Level	Sublevels
Realm	Duplodnaviria, Monodnaviria, **Riboviria**, Varidnaviria
Kingdom	**Orthornavirae**, Pararnavirae
Phylum	Duplornaviricota, Kitrinoviricota, Lenarviricota, Negarnaviricota, **Pisuviricota**
Class	Duplopiviricetes , **Pisoniviricetes**, Stelpaviricetes
Order	**Nidovirales**, Picornavirales, Sobelivirales

Taxonomic Level	Sublevels
Suborder	Arnidovirineae, **Cornidovirineae**, Mesnidovirineae, Monidovirineae, Nanidovirineae, Ronidovirineae, Tornidovirineae
Family*	**Coronaviridae**
Subfamily	Orthocoronavirinae, Torovirinae, **Coronavirinae**
Genus	Alphacoronavirus, **Betacoronavirus**, Deltacoronavirus, Gammacoronavirus
Subgenus	Embecovirus, Merbecovirus, Nobecovirus, **Sarbecovirus**

[3] https://talk.ictvonline.org/taxonomy/

Features used for the Machine Learning (ML) Model

Discrete Fourier Transform[2] (DFT)

$$F_i(k) = \sum_{j=0}^{p-1} f\left(S_i(j)\right) \cdot e^{(-2\pi i/p)kj}$$

- Finds the digital frequencies associated with numbers in a finite numeric sequence
- Prior study used the average magnitude of the Discrete Fourier transform for feature creation.

Shannon's Entropy

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Finds the measure of the intrinsic uncertainty embedded within a sequence
- Based on the concept that all systems have a tendency towards disorder

[2] Randhawa G, Soltysiak M, Roz HE, de Souza CPE, Hill KA, Kari L, Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study, PLOS One, 2020

Conversion Rules for Genomic Digitization

Illustration with the chosen conversion rule[2]: Purine Pyrimidine (PP)

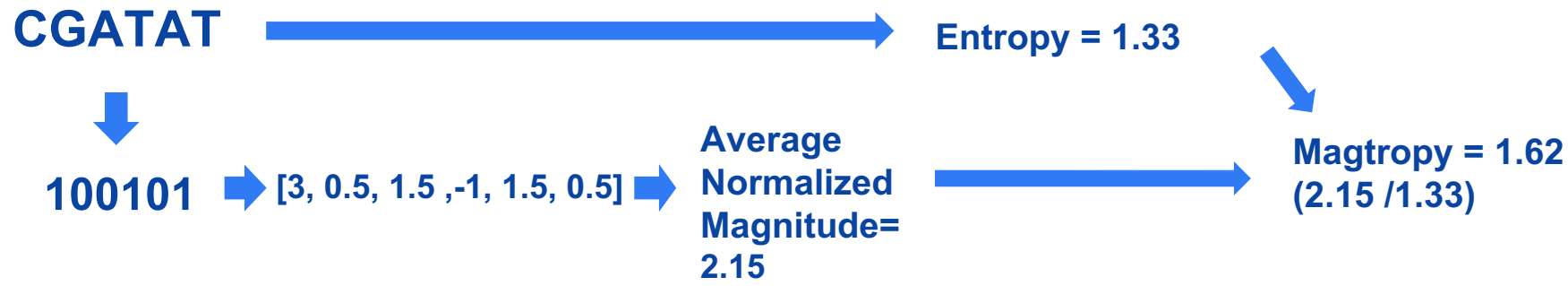
CAGGTCAT…. = 10001101….

Conversion Rule	A	T	C	G
Purine Pyrimidine (PP)	0	1	1	0
EIIP	0.13	0.14	0.15	0.08
Just A	1	0	0	0
Paired Numeric	1	1	-1	-1
Real	1.5	-1.5	0.5	-0.5
Integer 1	1	0	2	3
Integer 2	2	1	3	4
Just C	0	0	1	0
Just T	0	1	0	0
Just G	0	0	0	1

[2] Randhawa G, Soltysiak M, Roz HE, de Souza CPE, Hill KA, Kari L, Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study, PLOS One,2020

Processing the Data for Machine Learning

Step 1	Step 2	Step 3	Step 4	Step 5
Convert genomic sequences to numbers with PP conversion rule	Find frequency distribution of each nucleotide using DFT	Find absolute value of DFT (magnitude), then find the average	Separately, find entropy of the sequence	Divide the average magnitude by the entropy to find "magtropy"

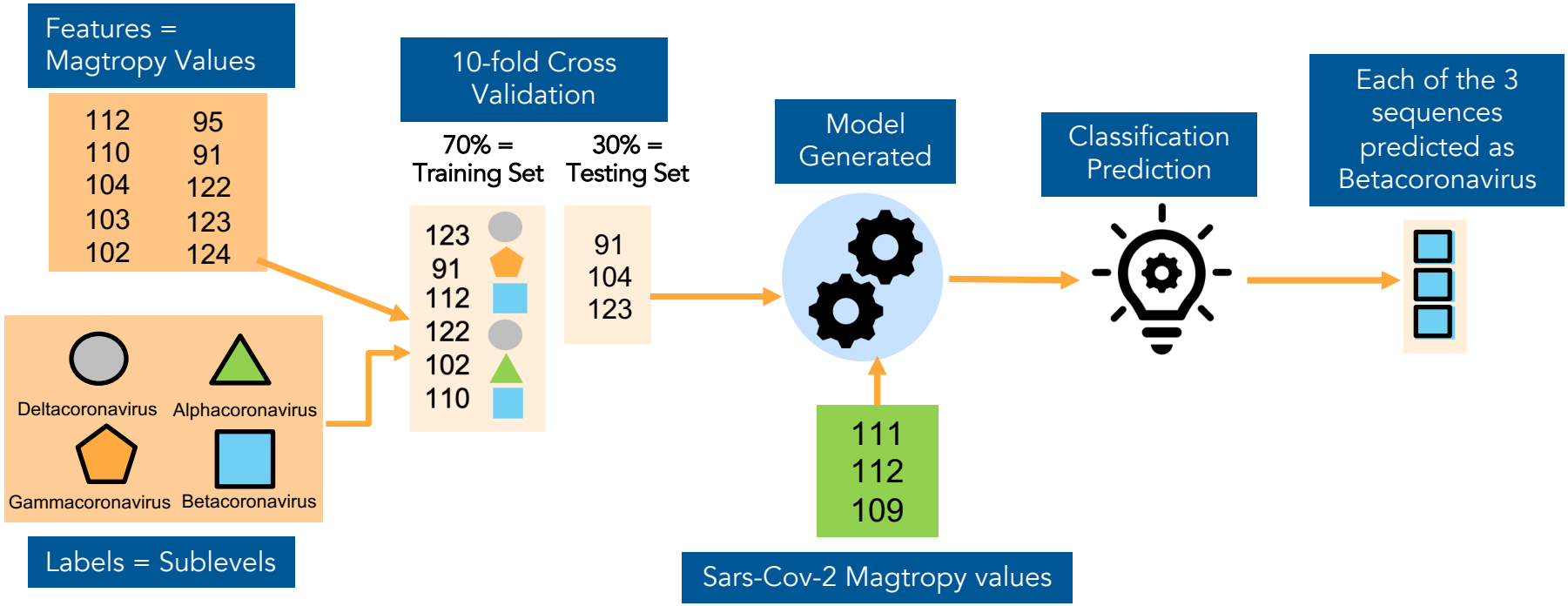


Raw vs. Processed Data for Genus level

Sublevel_Seq#	Sequence
Betacoronavirus_1	ATCGCGAGA….
Betacoronavirus_2	ATCGGGTCG….
Alphacoronavirus_1	GATGCTGTA…...
Alphacoronavirus_2	GAGTCTCTA…..
Gammacoronavirus_1	AGGCCAAAT…...
Gammacoronavirus_2	AGGTCAAAT…...
Deltacoronavirus_1	CCGGTAATA...
Deltacoronavirus_2	CAGGTAAAC...

Sublevel	Magtropy Value
Betacoronavirus	111.59
Betacoronavirus	110.72
Alphacoronavirus	103.75
Alphacoronavirus	102.98
Gammacoronavirus	95.88
Gammacoronavirus	90.74
Deltacoronavirus	121.78
Deltacoronavirus	125.87

Machine Learning Overview



Machine Learning Workflow

Sequences used for data

50 genomic sequences chosen at random from each sublevel

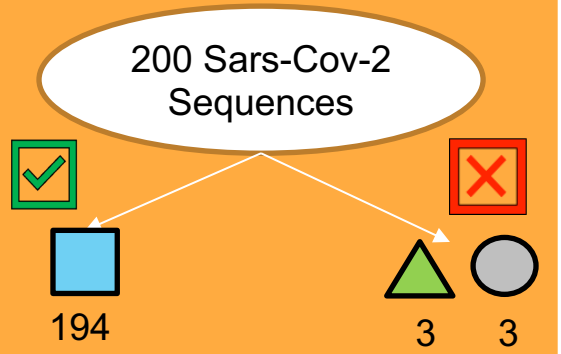
200 total

Classification Type:

Multi-class problem

One vs. rest ML technique

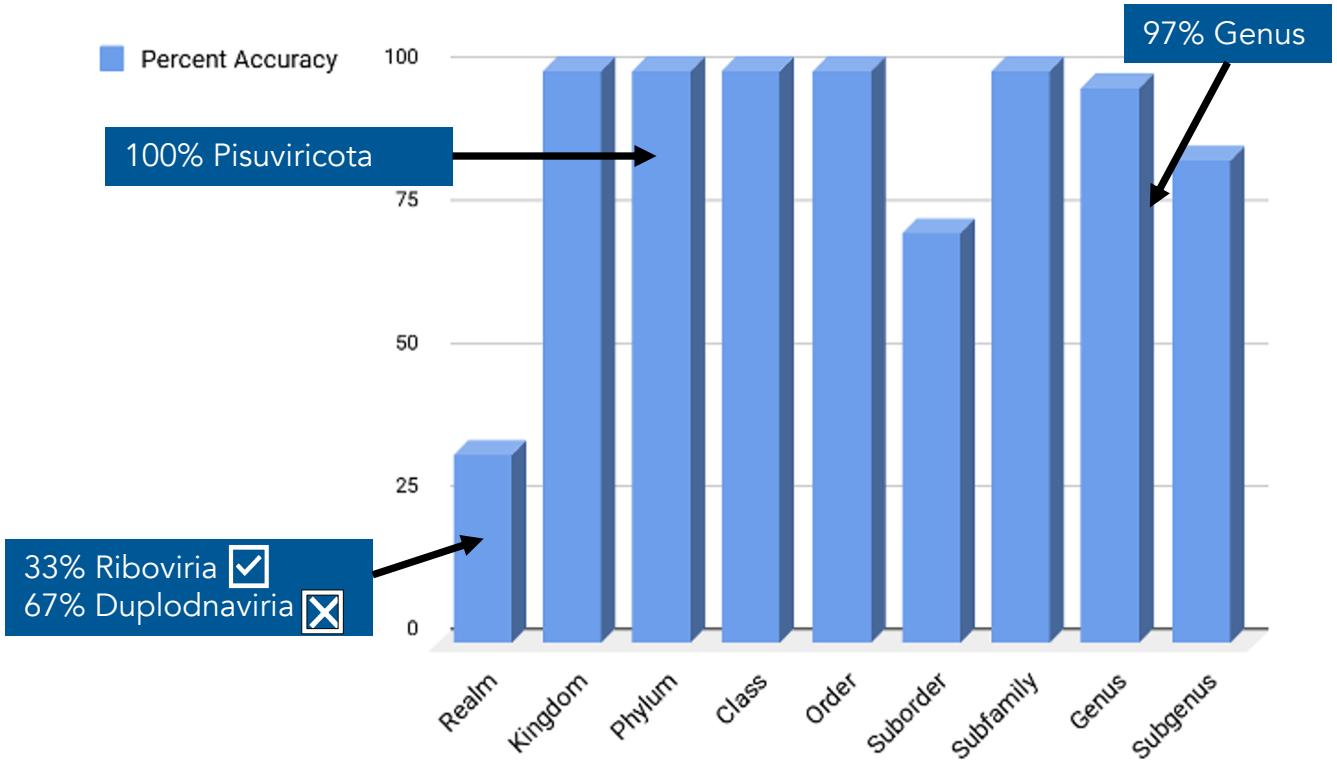
Optimized ML Metric: Accuracy



$$\text{Accuracy} = \frac{194}{200} = 97.0\%$$

Results with Purine Pyrimidine Conversion Rule

- 87.3% mean classification accuracy
- 2.5% accuracy in Phylum level with entropy alone, 100% with Magtropy
- Consistently best performing Classifiers: Extreme Gradient Boost, Decision Tree



Discussion

- Though DFT and Shannon's Entropy applied as distinct features in the ML model did not correctly classify Sars-Cov-2, combining them yielded a feature with substantially greater predictive power.
- Removing the subgenus and realm taxonomic levels increases mean classification accuracy to 95.5%
- Magtropy can be applied to further genomic classification studies.
- The methods developed are general enough to be applicable to genomic sequences from any organism.

Acknowledgements

A special thanks to the City of Hope Center for Informatics for supporting the bioinformatics summer internship program.

- Thank you to Mrs. Gallardo and the Brahma Tech Academy for constantly challenging me to put my best foot forward and believing in me.
- Thank you to Diamond Bar High School and Mr. Kevin Patterson for continuous guidance and support throughout the high school years.
- Thank you to my parents for encouraging me in all my endeavors!