UPPSALA UNIVERSITET

RESEARCH ARTICLE Open Access

A complete map of potential pathogenicity markers of avian influenza virus subtype H5 predicted from 11 expressed proteins

Zeeshan Khaliq[1], Mikael Leijon[2,3], Sándor Belák[3,4] and Jan Komorowski[1,5*]

Investigating potential virulence markers in the Nonstructural protein 1 of Avian Influenza Virus

Knowledge-based Systems for Bioinformatics

1MB465

Miltiadis Kesidis

Agapi Eleni Simaiaki

26/02/2024

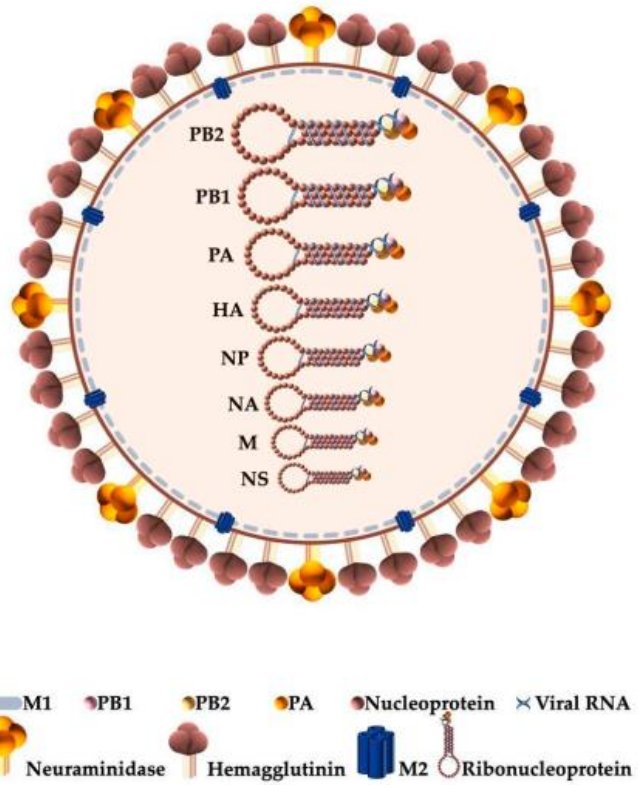
Introduction

Avian Influenza Viruses:

- Attacks human respiratory tract
- 10% of world population
- 290,000 - 650,000 annual deaths (WHO)
- High rates of mutation and recurrent genetic assortment
- Increased likelihood of pandemic

Structure:

- RNA virus
- 8 single-stranded RNA segments
- Replicates in host's nucleous



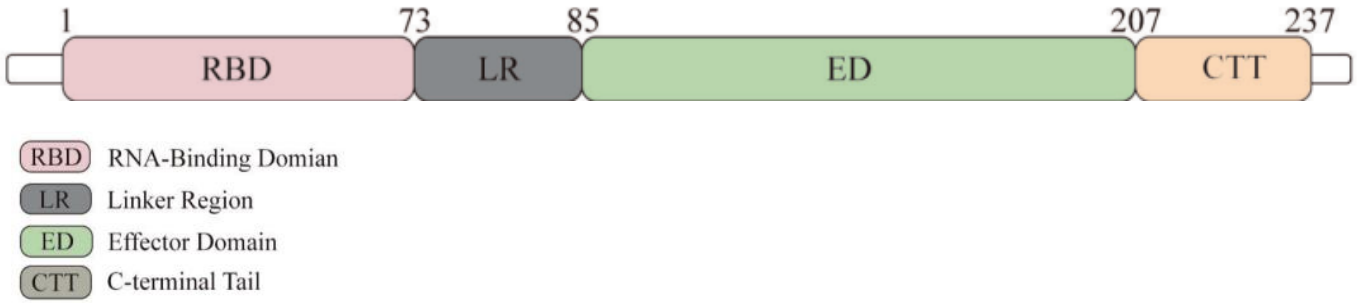
Z. Ji, X. Wang, et. Al., 2021

PB1—polymerase basic protein 1
PB2—polymerase basic wrapped in viral RNA
PA—polymerase acidic protein
NP – Nucleoprotein
NA – Neuraminidase
HA – Hemagglutinin
M – Matrix protein (M1 – M2)
NS – Nonstructural protein (NS1 – NS2/NEP)

Introduction

Nonstructural protein 1 (NS1):

- 230-237 amino acids
- Counteracts host's natural immune support
- Blocks interferon production
- Aids viral reproduction



Z. Ji, X. Wang, et. al., 2021

Goals of study:

- Pathogenicity associated with HA
- Monobasic cleavage site -> LP
- Insertions in cleavage site -> HP
- Yet not enough
- **Search for additional virulance factors in NS1**

Methods - Feature Selection

Dataset (328x251):

- 328 objects – Influenza A, H5 subtype strains
- 250 features / aligned amino acid positions
- 1 decision class column
- 1-> High pathogenicity (HP)
- 0 -> Low pathogenecity (LP)
- Balanced dataset

```
library(rmcfs)

mcsf_fts <- mcfs(Pathogenicity~., data, projections=500,
                 projectionSize=50, splits=5, splitSetSize = 40,
                 cutoffPermutations = 20, threadsNumber = 8)
```

Monte Carlo Feature selection – mcfs:

- Tested combination:
  - **500 projections** – 25/**50**/75 set size
  - 750 projections – 25/50/75 set size
  - 1000 projections – 25/50/75 proj set size
- Splits: 5
- Training split: 80/20
- Cutt off permutation: 20

R package:

- rmcfs::v85i12 Draminski & Koronacki (2018): rmcfs paper (Journal of Statistical Software)

UPPSALA UNIVERSITET

Methods – R.Rosetta

Running Rosetta:

- Input: most significant features (mcfs)
- Reducts algorithms tested:
  - **Johnson (Acc:0.941 / AUC:0.96 / 62 rules)**
  - Genetic (Acc:0.938 / AUC: 0.97 / 734 rules)
- Multiple testing Correction: **Bonferroni**

```
dataJohnson <- rosetta(mcsf_fts$data,reducer = "Johnson", roc = TRUE,
                       discrete = TRUE, clroc = "1")

dataGenetic <- rosetta(mcsf_fts$data, reducer = "Genetic", roc = TRUE,
                       discrete = TRUE, clroc = "1")
```

Extraction of most significant rules:

- P-value <0.05 - 16 rules
- Based on Accuracy - 0.92 & Coverage – 0.43

R packages:

- R.ROSETTA: an interpretable machine learning framework
- VisuNet: an interactive tool for network visualization of complex rule-based classifiers.

Results - R.Rosetta

Top rules (p-value < 0.05):

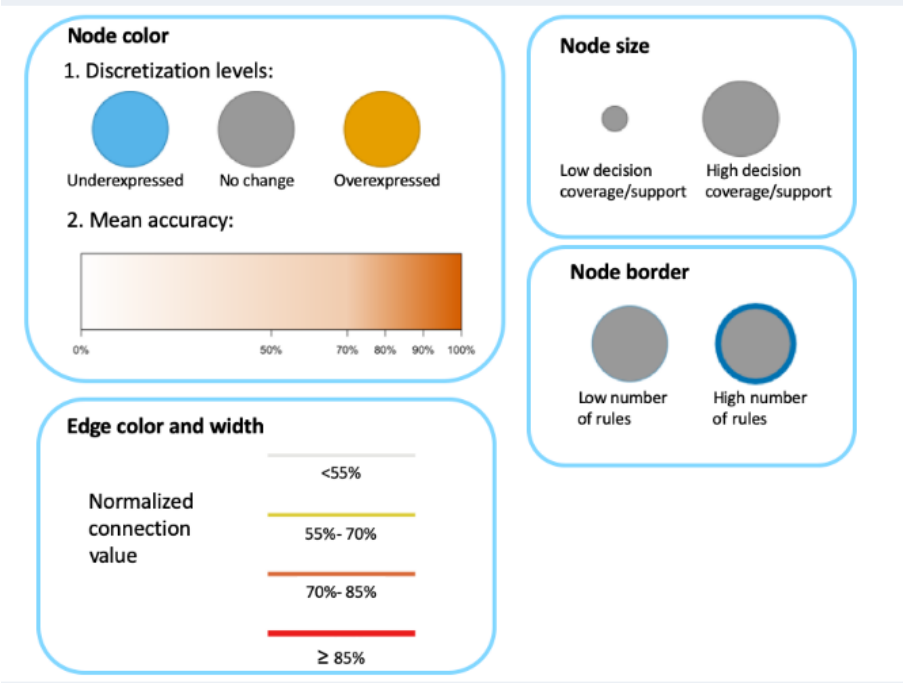
High pathogenicity :

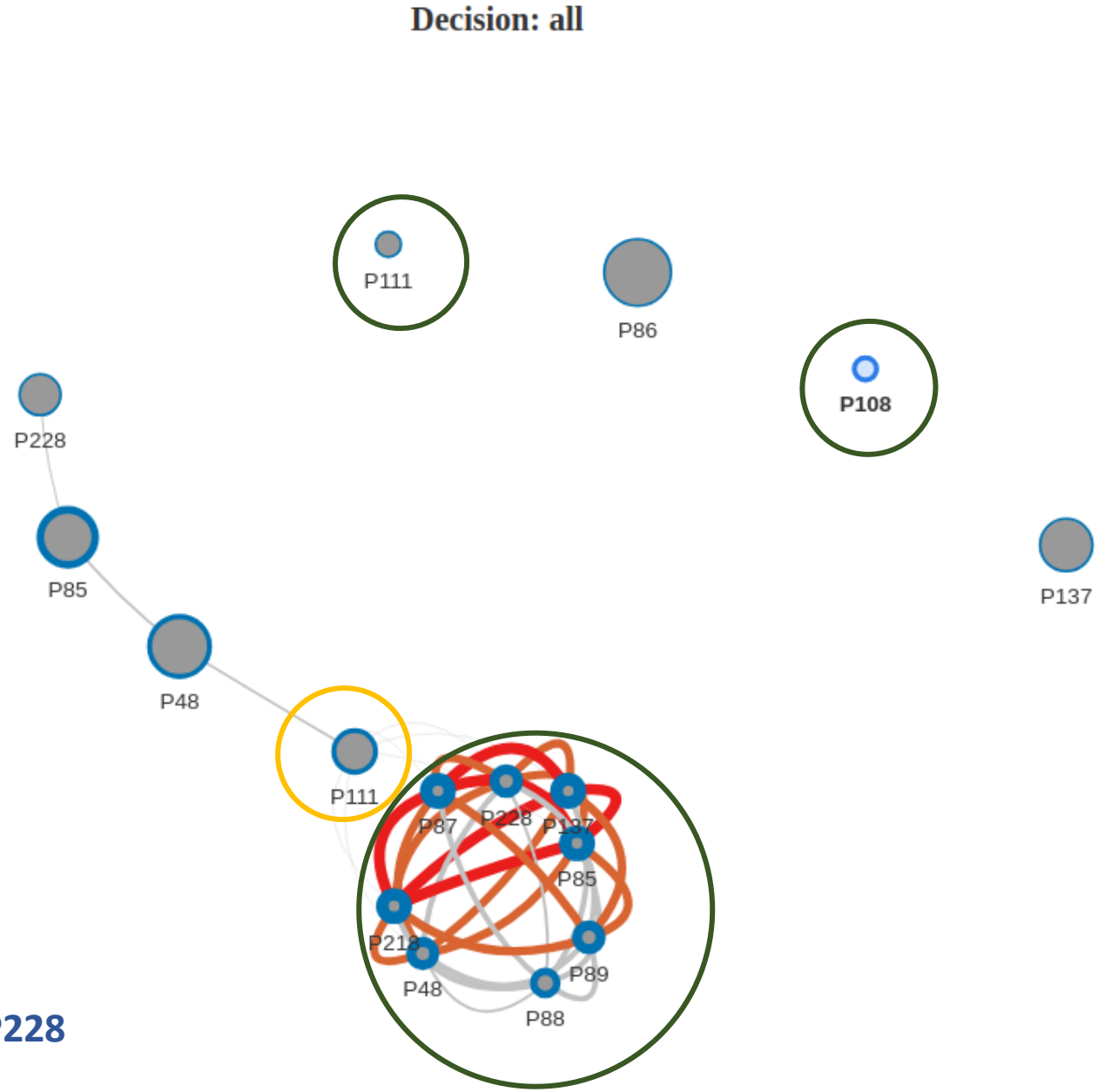
	rule	length	accuracy	support	pValue
1	IF P86(?) THEN 1	1	1.00000	135	6.071491e-62
2	IF P85(?) THEN 1	1	0.99298	128	1.760143e-54
3	IF P48(N) AND P111(D) THEN 1	2	0.99291	122	3.064380e-50
4	IF P48(N) AND P85(?) THEN 1	2	1.00000	118	1.506834e-49
5	IF P137(T) THEN 1	1	0.98148	106	1.183716e-38
6	IF P85(?) AND P228(N) THEN 1	2	1.00000	92	1.608758e-34

Low pathogenicity:

	rule	length	accuracy	support	pValue
1	IF P108(I) THEN 0	1	1.00000	66	2.176941e-22
2	IF P48(S) AND P85(T) AND P87(A) AND P89(V) AND P137(N) AND P218(N) AND P228(K) THEN 0	7	0.95890	70	6.665395e-20
3	IF P111(E) THEN 0	1	0.95026	70	7.852161e-19
4	IF P85(T) AND P87(A) AND P88(S) AND P89(V) AND P137(N) AND P218(N) AND P228(K) THEN 0	7	0.94444	68	5.033903e-18
5	IF P48(S) AND P87(A) AND P89(V) AND P111(D) AND P128(R) AND P137(N) AND P218(N) AND P228(K) THEN 0	8	0.95522	64	1.791682e-17
6	IF P48(S) AND P85(T) AND P87(A) AND P88(S) AND P89(V) AND P137(N) AND P218(N) AND P228(K) THEN 0	8	0.94286	66	3.134982e-17
7	IF P48(S) AND P85(T) AND P87(A) AND P89(V) AND P111(D) AND P137(N) AND P228(K) THEN 0	7	0.94030	63	4.619977e-16
8	IF P48(S) AND P85(T) AND P87(A) AND P89(V) AND P111(D) AND P137(N) AND P218(N) AND P228(K) THEN 0	8	0.93539	63	4.619977e-16
9	IF P48(S) AND P85(T) AND P87(A) AND P88(S) AND P89(V) AND P137(N) AND P218(N) THEN 0	7	0.94030	63	4.619977e-16
10	IF P48(S) AND P85(T) AND P87(A) AND P111(D) AND P137(N) AND P218(N) AND P228(K) THEN 0	7	0.92537	62	9.239125e-15

Results - VisuNet

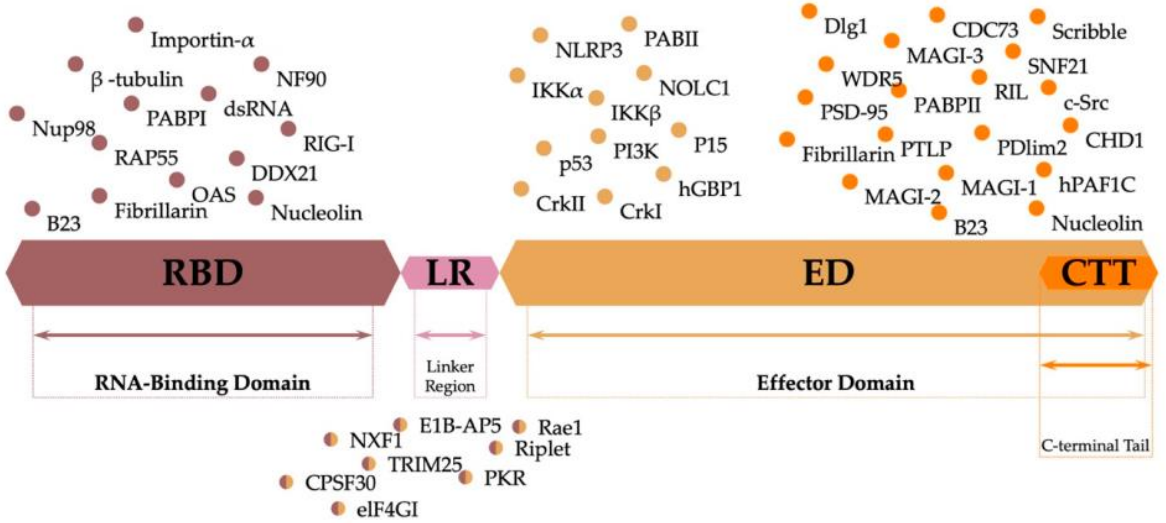




Low pathogenicity

Part of both classes: P48, P85, P111, P137, P228

Discussion



Rosário-Ferreira, N., Preto, A. J.*et. al., 2020*

Exploring common features :

- P48: HP-> Asn vs LP -> Ser : Associated with dsRNA and transport of virus to nucleus
- P85: HP -> GAP vs LP-> Thr: Inhibit host innate immune response
- P111: HP -> Asp vs LP -> Glu OR Asp with P48 Ser: Inhibit innate immune response
- P137: HP -> Thr vs LP -> Asn: nuclear export signal (NES) at residues 138-147
- P228: HP -> Asn vs LP -> Lys : Allows for Aketylation / virus polymerase activity and replication

Discussion

Interesting observations:

- Amino Acid changes of similar functionality
- Most changes in Effector domain

Further Investigation:

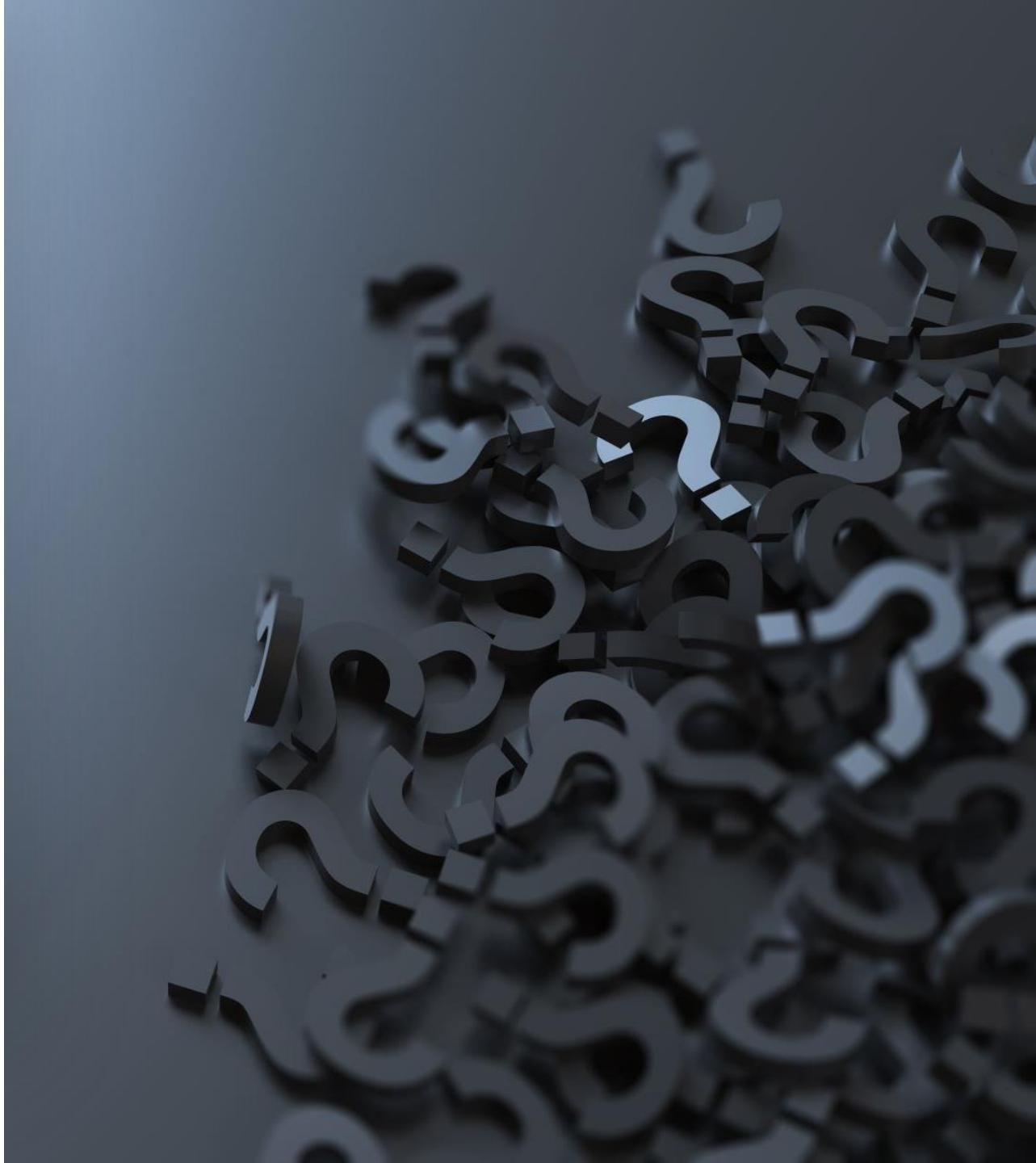
- Investigate further Genetic algorithm
- Check more flexible multiple testing correction for Johnson
- Validation of our model with unseen data / Update dataset

UPPSALA UNIVERSITET

Thank you!

References:

- Dramiński, M. and Koronacki, J. 2018. rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. Journal of Statistical Software. 85, 12 (Jul. 2018), 1–28. doiI:https://doi.org/10.18637/jss.v085.i12
- Dramiński, M., Rada-iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., & Komorowski, J. 2008. Monte Carlo feature selection for supervised classification. Bioinformatics, 24(1), 110–117. https://doi.org/10.1093/bioinformatics/btm486
- Garbulowski M, Diamanti K, Smolińska K, et al. R.ROSETTA: an interpretable machine learning framework. *BMC Bioinformatics*. 2021;22(1):110. Published 2021 Mar 6. doi:10.1186/s12859-021-04049-z
- Ji ZX, Wang XQ, Liu XF. NS1: A Key Protein in the "Game" Between Influenza A Virus and Host in Innate Immunity. *Front Cell Infect Microbiol*. 2021;11:670177. Published 2021 Jul 13. doi:10.3389/fcimb.2021.670177
- Khaliq Z, Leijon M, Belák S, Komorowski J. A complete map of potential pathogenicity markers of avian influenza virus subtype H5 predicted from 11 expressed proteins. *BMC Microbiol*. 2015;15:128. Published 2015 Jun 26. doi:10.1186/s12866-015-0465-x
- Rosário-Ferreira N, Preto AJ, Melo R, Moreira IS, Brito RMM. The Central Role of Non-Structural Protein 1 (NS1) in Influenza Biology and Infection. *Int J Mol Sci*. 2020;21(4):1511. Published 2020 Feb 22. doi:10.3390/ijms2104151



UPPSALA UNIVERSITET

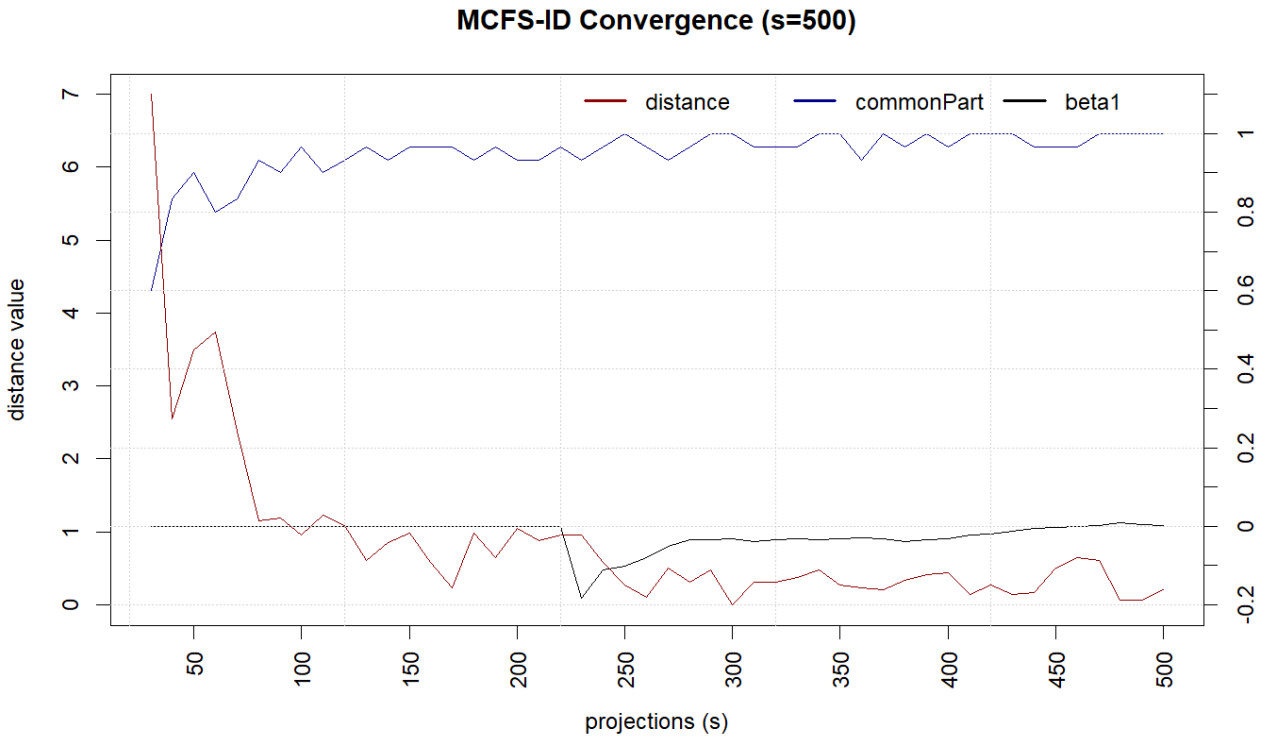
Influenza A – RNA segments & protein function

RNA Segment	Protein(s) Coded	Function [20,31,35]	Structural Data [20,31,35,36]
1	**PB2** 759 aas	Located in the nucleus of infected cells; Signals the viral polymerase passage to the host's nucleus; Enhances the formation of the cap structures necessary for viral messenger RNA (mRNA) transcription; Located in the mitochondria of infected cells [37]; Inhibits Interferon-β; Helps determine host range.	The three proteins, PB2 (polymerase basic protein 2), PB1 (polymerase basic protein 1) and PA (polymerase acidic protein), form the viral RNA polymerase, responsible for viral RNA transcription and replication.
2	**PB1** 757 aas	Responsible for the elongation of the primed nascent viral mRNA; Located in the nucleus of infected cells; Enhances the association of the 3 subunits of the RNA polymerase complex.	
3	**PA** 716 aas	Functions still unknown, but evidence points to helicase-like functions; Important for viral transcription; Assembly of the polymerase complex.	

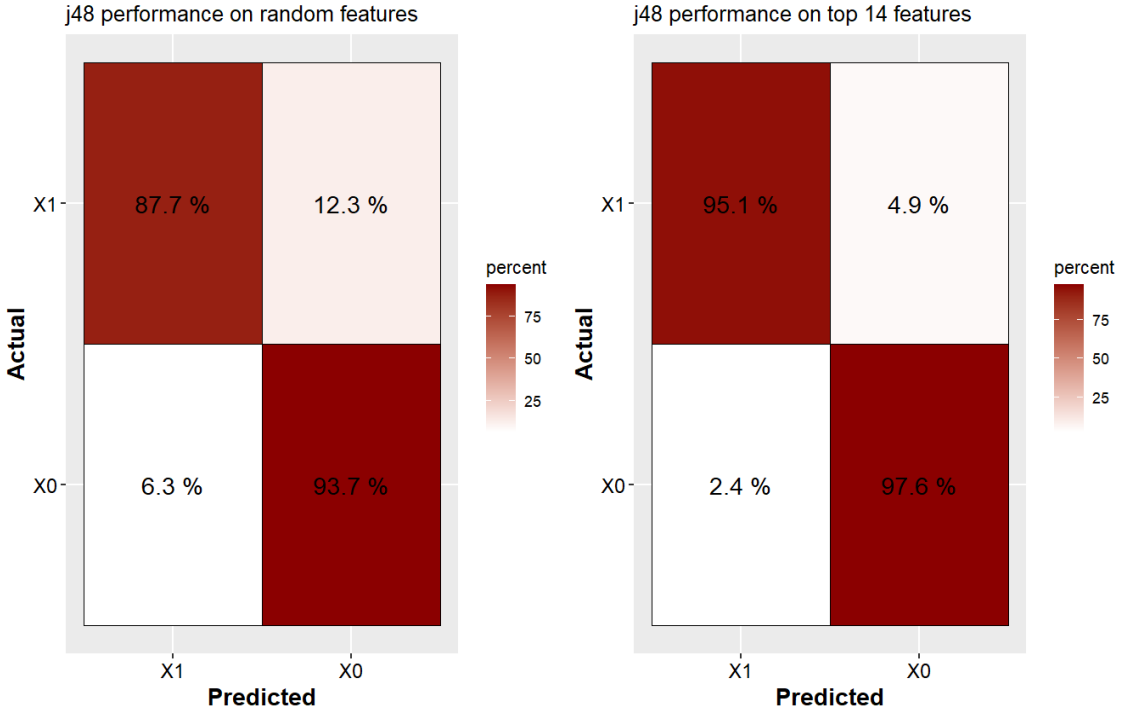
RNA Segment	Protein(s) Coded	Function [20,31,35]	Structural Data [20,31,35,36]
4	**HA** 550 aas	Attaches the virions to the sialic acid (SA) moieties of the host's receptors; Around 30% variation between subtypes.	Hemagglutinin (HA) is a homotrimeric integral cylinder-like membrane glycoprotein on the virus surface; 4 antigenic sites with direct impact on virulence and pathogenicity of the virus.
5	**NP** 498 aas	Binds non-specifically to single-stranded RNA (ssRNA); Encapsidates viral RNA; Helps recruiting RNA polymerase for synthesis of viral positive-sense RNA (cRNA); Related to host range.	Nucleoprotein (NP) is a 56 kDa basic protein; RNA-binding protein; Structural unit of RNPs; Forms oligomers stabilized by vRNA.
6	**NA** 470 aas	Unnecessary for virus replication; Required for budding of newly formed viral particles from surface of infected cells; Facilitates virus movement to the target cell by cleavage of sialic acids from respiratory tract mucins; Helps the release of virions from infected cells.	Neuraminidase (NA) is a homotetrameric integral membrane glycoprotein with 4 structural domains; Antigenic sites help circumvent the immune responses aiding on the virulence and pathogenicity of the virus.
7	**M1** 252 aas	Membrane-binding and RNA-binding protein; Forms a coat inside the viral envelope; Determines virion's shape; Interacts with vRNP and other cytoplasmic domains of integral membrane proteins; Increases vRNPs export and decreases import; Helps assembly and budding of virions.	Matrix protein (M1) formed by a globular N-terminal domain and a flexible C-terminal tail; Oligomerization state and binding to lipid bilayer are highly dependent on pH.
	M2 97 aas	Vital for viral replication; Forms proton channel in virus envelope; Lowers the pH inside the viral particle to promote uncoating of RNPs; Modulates Golgi's pH; Helps to stabilize HA's native conformation during virus assembly.	Matrix-2 protein (M2) is a 97-residue single-pass membrane protein; Three segments: N-terminal outward segment, transmembrane (TM) helix, and C-terminal inward segment; TM helices from 4 subunits pack to form proton-channel; Highly conserved His37 and Trp41 residues.
8	**NS1** 230 aas	NS1 acts as a promoter of viral replication and an inhibitor of the host's immune response; Present in the cytoplasm and nucleus of the host cell.	Non-structural protein 1 (NS1) has two structural domains—RNA-binding domain (RBD) and the effector domain (ED)—connected by a short linker (LR), and a disordered C-terminal tail (CTT).
	NEP/NS2 121 aas	Promotes viral RNA replication; Regulates vRNP's export from the nucleus to the cytoplasm; RNA nuclear export; Interacts with the viral matrix M1 protein.	Nuclear Export Protein (NEP) has a protease-sensitive N-terminal domain (residues 1–53) and a protease-resistant C-terminal domain (residues 54–121) mostly formed by a helical hairpin.

MCFS - Feature Selection parameter selection

- **500 projections** – 50 projection set size
- Splits: 5
- Training split: 80/20
- Cutt off permutation: 20



Distance plot - Interdependency discovery – convergence diagnostics of the algorithm



Confusion matrix obtained on all 500*50 trees

UPPSALA UNIVERSITET

Rosetta Reduct Validation - Johnson