Flinders
UNIVERSITY


# Identifying mutation positions in all segments of influenza genome enables better differentiation between pandemic and seasonal strains

Fatemeh Kargarfard, Ashkan Sami, Farhid Hemmatzadeh, Esmaeil Ebrahimie

PII: S0378-1119(19)30058-7
DOI: https://doi.org/10.1016/j.gene.2019.01.014
Reference: GENE 43536

To appear in: *Gene*

Received date: 24 March 2018
Revised date: 29 December 2018
Accepted date: 17 January 2019

Please cite this article as: F. Kargarfard, A. Sami, F. Hemmatzadeh, et al., Identifying mutation positions in all segments of influenza genome enables better differentiation between pandemic and seasonal strains, Gene, https://doi.org/10.1016/j.gene.2019.01.014

# Identifying mutation positions in all segments of influenza genome enables better differentiation between pandemic and seasonal strains

Fatemeh Kargarfard[1,2], Ashkan Sami[1]*, Farhid Hemmatzadeh[2], Esmaeil Ebrahimie[2,3,4,5]*

[1]Faculty of Engineering and IT, University of Technology Sydney, New South Wales, Australia. [2]Department of Computer Science and Engineering, School of Electrical Engineering and Computer, Shiraz University, Shiraz, Iran. [2]School of Animal and Veterinary Sciences, The University of Adelaide, Adelaide, Australia. [3]Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, Australia. [4]School of Information Technology and Mathematical Sciences, Division of Information Technology Engineering & Environment, University of South Australia, Adelaide, Australia. [5]School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, Australia.

*Corresponding author

Dr. Esmaeil Ebrahimie

Adelaide Medical School

Faculty of Health and Medical Sciences

The University of Adelaide,

Adelaide, Australia

Ph : +61 8 831 37874

Email: esmaeil.ebrahimie@adelaide.edu.au

## Abstract

Influenza has a negative sense, single-stranded, segmented RNA. In the context of pandemic influenza research, most studies have focused on variations in the surface proteins (Hemagglutinin and Neuraminidase). However, new findings suggest that all internal and external proteins of influenza viruse can contribute in pandemic emergence, pathogenicity and increasing host range. The occurrence of the 2009 influenza pandemic and the availability of many external and internal segments of pandemic and non-pandemic sequences offer a unique opportunity to evaluate the performance of machine learning models in discrimination of pandemic from seasonal sequences using mutation positions in all segments. In this study, we hypothesized that identifying mutation positions in all segments (proteins) encoded by the influenza genome would enable pandemic and seasonal strains to be more reliably distinguished. In a large scale study, we applied a range of data mining techniques to all segments of influenza for rule discovery and discrimination of pandemic from seasonal strains. CBA (classification based on association rule mining), Ripper and Decision tree algorithms were utilized to extract association rules among mutations. CBA outperformed the other models. Our approach could discriminate pandemic sequences from seasonal ones with more than 95% accuracy for PA and NP, 99.33% accuracy for NA and 100% accuracy, precision, specificity and sensitivity (recall) for M1, M2, PB1, NS1, and NS2. The values of precision, specificity, and sensitivity were more than 90% for other segments except PB2. If sequences of all segments of one strain were available, the accuracy of discrimination of pandemic strains was 100%. General rules extracted by rule base classification approaches, such as M1-V147I, NP-N334H, NS1-V112I, and PB1-L364I, were able to detect pandemic sequences with high accuracy. We observed that mutations on internal proteins of influenza can contribute in distinguishing the pandemic viruses, similar to the external ones.

**Key Words**: Association rule mining; CBA; Expert system; Hot spots; Ripper algorithm; Pandemic Influenza

## Introduction

Influenza A belongs to the Orthomyxoviridae family with a negative sense, single-stranded, segmented RNA. This virus has 8 segments: HA (hemagglutinin), NA (neuraminidase), NP (nucleoprotein), M (two matrix proteins, M1 and M2), NS (two distinct non-structural proteins, NS1 and NS2), PA (RNA polymerase and PA-X), PB1 (RNA polymerase and PB1-F2 protein), and PB2 (RNA polymerase) (Horimoto and Kawaoka, 2005).

Even a small number of mutations in the hemagglutinin gene of H1N1 influenza has the potential to change antigenic characteristics and cause a significant reduction in the immunity of human populations and vaccine efficiency (Strengell et al., 2011; Ebrahimi et al., 2014a; Ebrahimie et al., 2015; Tarigan et al., 2018). Similar to hemagglutinin and neuraminidase changes, mutation/re-assortment can also occur in other viral proteins including internal segments. The co-occurrence of mutations on external and internal segments, such as HA-E391K with PB2-K340N, has been reported (Maurer-Stroh et al., 2009a). Also, the co-occurrence of HA-E391K, HA-D114N, PB1-R563K, and PA-V14I in a Spanish strain is observed (Maurer-Stroh et al., 2009a). Detecting mutations in each influenza protein sequence, either external or internal, and finding the combination/patten of mutations is an important step in discrimination of pandemic sequences from seasonal ones.

For pattern recognition in influenzas sequences, many studies have focused on visual alignment of a regional subset (maximum 200) of sequences and application of multivariate techniques such as clustering (Ebrahimi et al., 2014b). Multivariate clustering methods (e.g., UPGMA and neighbor-joining) are routinely used to classify lineages into different categories. However, clustering methods ignore the quality of mutations, and all points have the same value in contributing to final classification and prediction, questioning their ability to determine mutational hot spots efficiently (Ebrahimi et al., 2014a). Immunological tests (such as ELISA or western blot) have also been used to detect decreased antigen-antibody responses following mutation in the key positions (Strengell et al., 2011; Hemmatzadeh et al., 2013; Hadifar et al., 2014; Hasan et al., 2016). Most of these studies are limited to one or two proteins (in particular HA and NA) and they ignore the effects of other proteins in pandemic occurrence (Ebrahimi et al., 2014b).

Sequencing a large number of influenza segments from the 2009 pandemic has provided a unique opportunity and a valuable source of data to examine the performance of various supervised machine learning models, such as CBA, in discrimination of pandemic influenza from the seasonal ones (Kargarfard et al., 2015).

Machine learning algorithms are the method of choice for better understanding of various phenomena, extracting implicit, actionable and previously unknown rules and providing prediction capabilities (Sivathayalan, 2009; Bakhtiarizadeh et al., 2014; Shekoofa et al., 2014; Mohammadi-Dehcheshmeh et al., 2018). The final aim of these data mining techniques is to extract knowledge (underlying rules) from a dataset and converting this knowledge into a perceptible format for further use (Jamali et al., 2016; Ebrahimie et al., 2018; Sharifi et al., 2018). The most popular method for discovering relations in a dataset is "association rule generation" (Ebrahimi et al., 2010; Kargarfard et al., 2015; Kargarfard et al., 2016). Human readable rules imply to data presented in a format which is readily interpreted by humans. Normally these rules follow the "IF… THEN" format. This representation of knowledge is the most appropriate manner for biologists and virologists to express their knowledge in finding significant genetic markers. For example, IF (HA-E391K) AND (HA-D114N) THEN (the mortality rate is high). IF… THEN rules structure is modular, relatively small, and informative. These rules can be applied as a basis for classification of instances (Daud and Corne, 2009). Associative classification methods are recent machine learning strategies to build a classifier based on rules that integrate classification with association rule mining. Some accurate and effective classifiers based on associative classification are: CBA (Classification Based on Associations) (Bing Liu, 1998), CMAR (Classification based on Multiple Association Rules) (Li et al., 2001), and CPAR (Classification based on Predictive Association Rules) (Yin and Han, 2003).

Sequencing a large number of influenza segments from the 2009 pandemic has provided a unique opportunity and a valuable source of data to examine the performance of various supervised machine learning models, such as CBA, in discrimination of pandemic influenza from the seasonal ones (Kargarfard et al., 2015). The main goal of this study was to identify the potential mutations associated with influenza pandemic occurrence by analysis of the whole viral genome/proteome, instead of analysis of 1 or 2 proteins. This study extends our recent findings regarding the HA segment (Kargarfard et al.,

2015) to all viral protein segments. This study provides the underlying knowledge for recognition of key mutations in the viral sequence and their co-occurrence interactions (based on association rules).

## Material and Methods

In this study, CBA, Ripper and Decision tree algorithms were utilized for extracting the association rules among mutations. Figure 1 displays an overview of steps of this research. Process implementation is discussed in detail in the following sections.

### Dataset

To select the 2009 pandemic sequences, the parameter "Include only pH1N1 proteins" was selected on Influenza Research Database (https://www.fludb.org/)(Squires et al., 2012). To download the seasonal sequences, the parameter "Exclude all pH1N1 proteins" was selected. File S1 - S10 include protein sequences. Table 1 represents more information about the dataset.

To validate the findings and to prevent the overfitting of the discovered rules, another dataset containing all segments of H1N1 was used. None of the sequences in this dataset was involved in extracting rules (unseen data). We named this new dataset "test data". The data was downloaded from Influenza Research Database (IRD). The HA nucleotide sequences which were used for extracting rules is presented at File S21.

To execute the rule based algorithms, a dataset was generated including all segments of influenza sequences. Dataset contained 10 proteins because each of $7^{th}$ or $8^{th}$ segment produces two proteins. In addition to protein sequences, nucleotide sequences were downloaded for HA segment. Only complete sequences were downloaded in this research. These sequences were separated into two groups: pandemic and seasonal. Pandemic sequences comprised of the 2009 flu pandemic. The data were downloaded from Influenza Research Database (IRD) which is a resource for the influenza virus research community to facilitate an understanding of the influenza virus (Squires et al., 2012) .

### Data Preparation

MUSCLE algorithm was used for multiple sequence alignment. MUSCLE stands for multiple sequence comparison by log-expectation. It is one of the most well-known multiple alignment software for protein and nucleotide sequences (Edgar, 2004). Commonly, MUSCLE gives higher average accuracy and better speed compared to the other multiple alignment tools such as CLUSTALW (Larkin et al., 2007) or T-Coffee (Notredame et al., 2000), by selecting the maximum amount of iterations and diagonal optimization. MUSCLE has three phases. At the end of each phase, the multiple alignment can be obtained and the algorithm can be terminated. Phase 1 is draft progressive. Phase 2 is improved progressive. The final phase (Phase 3) executes iterative improvement based on a variant of tree-dependent restricted partitioning (Attaluri et al., 2009).

The variables 'maximum iteration' and 'maximum memory in MB' at MUSCLE were set to 2 and 3000 MB, respectively. Because of the large size of dataset, only the first two iterations of the algorithm were performed. After sequence alignment, data were stored in relational table; it has a set of attributes. Features or attributes represent the nucleotide or amino acid at each position in a sequence (for example Att12 means $12^{th}$ position of sequence). In the case of CBA tool, data was converted into C4.5 format (*.data, * .names files).

**Rule generation**

For rule extraction in detecting pandemic sequences, first, we applied CBA, Ripper and C4.5 (decision tree) algorithm on different protein segments. "RapidMiner" software (2015) was used for running Ripper and C4.5. To obtain generalized and accurate rules, we assigned minimum support to be 10% and the minimum confidence to 90%.

Rule based classification algorithms generate many rules which some of them may not be appropriate for our goal. We selected the rules according to the following three indications: (1) Coverage of a rule (support) to be more than 10%, (2) Accuracy of a rule (confidence) to be more than 90%, (3) Length or number of descriptors was set two as the maximum length of the rules. In the current study, among all the generated rules, only the rules with high support and confidence were selected.

**Decision tree**

A decision tree is an expressive representation intended for classifying instances. The purpose is to construct a model which predicts the value of a target variable according to numerous input

parameters. In these tree structures, class labels are represented by leaves and branches symbolize conjunctions of features which result in those class labels. A tree is usually "learned" through dividing the original set into subsets according to an attribute value test. This process is replicated upon every taken subset in the recursive approach called recursive partitioning. The recursion is finished when the subset of a node has all the identical value of the target feature, or when more division do not add more value to the predictions (Rokach, 2008).

Many specific Decision-tree algorithms exist. Notable ones are: ID3, C4.5, and CHAID (Kass, 1980; Ebrahimi et al., 2011). We used C4.5 for extracting rules. At every node of the tree, C4.5 selects the attribute of the data that most properly divides its set of instances into subsets enriched in one class or the other. The division measure is the normalized information gain (difference in entropy). The feature with the highest normalized information gain is taken to build the decision. This process is replicated on the smaller subset (Quinlan, 1993). (Sharifi et al., 2018)

**Ripper algorithm**

In this study, we applied a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is an optimized version of IREP. The algorithm is briefly described as follows:

RIPPER consists of two phases. In the first phase, a rule set was built by repeatedly adding rules to an empty rule set until no positive examples (pandemic sequences) exist, or the error rate >= 50%. Rules were formed by adding antecedents greedily (or conditions) to the rule until the rule was perfect (i.e. 100% accurate). After a rule set was constructed, each rule was pruned incrementally and let the pruning of any final sequences of the antecedents. In the second phase, an optimization was performed on the rule set in order to decrease its size and improve its fitness to the training data. (William, 1995). More explanation about the Ripper algorithm is provided in Supplementary InfoFile 1.

**CBA Algorithm**

CBA is an integrative algorithm which has the power of both classification and association rule. This integration was performed by mining class association rules (CARs). The following definitions for association rules were used in this study:

1) Let D be a relational table with n attributes.
2) Assume I be the set of all items in D, and Y be the set of class labels.
3) A class association rule (CAR) is an implication of the form: $X \rightarrow y$, where $X \subseteq I$, and $y \in Y$.
4) The rule $X \rightarrow Y$ holds in the transaction set D with confidence c, if c% of transactions in D that contains X also contains Y
5) The rule $X \rightarrow Y$ has supports in the transaction set D if s% of transactions in D contains $X \cup Y$ given a set of transactions D (In other words, Support states how frequent the items appear in the database and confidence represents the number of times which the statements have been found to be true)
6) ruleitem: <condset, y >, representing the rule: condset → y, where *condset* is a set of items, $y \epsilon Y$ is a class label (Agrawal and Srikant, 1994).

CBA has 2 parts:

1) A rule generator (called CBA-RG) is defined based on algorithm Apriori for finding association rules. The CBA-RG algorithm generates all frequent *ruleitems* by making multiple passes over the data. (Agrawal and Srikant, 1994)
2) A classifier builder (called CBA-CB). The CBA-CB algorithm builds a classifier using CARs. To produce the best classifier out of the whole set of rules, a minimum number of rule sets would be selected to cover the training dataset and minimize the lowest error rate (Bing Liu, 1998).

In this study, I is a set of nucleotides or amino acids. For protein sequences, I includes 20 members such as A, R, N, D, C, Q. Every protein sequence represents a transaction (T). T is a subset of I. All the sequences together construct the set. D Class labels ($c_i$) are either pandemic or seasonal. More explanation about the CBA algorithm is provided in Supplementary InfoFile 1.

## Results

### Discovered rules for each segment

CBA outperformed the other models. Based on the rules generated by CBA, nucleotide sequences of HA were classified with 99.99% accuracy, and protein sequences with 99.98% accuracy. M1 sequences were classified with 99.77% accuracy based on the extracted rules. Table 2 represents the extracted rules on nucleotide sequences of HA segment and Table 3 represents the extracted rules on M1 protein sequences by Ripper algorithm.

The extracted rules and their corresponding supports are provided in supplementary files (please see Table S1 – S9). The accuracy of NA sequences was 99.94%; 99.73% for M2 sequences, 99.57% for NP sequences, 97.58% for PA sequences, 99.88% for PB1 sequences, 82.54% for PB2 sequences, 97.27% for NS1 sequences, and 98.77% for NS2 sequences. The extracted rules are visualized at Figure S1-S7.

### Discovered rules governing 2009 pandemic occurrence

Rule 1 in Table 2 states that in 67.39 % of nucleotide sequences of HA, when $260^{th}$ position is not 'T' (Thymine), the sequences is pandemic. Rule 2 states that when this position is 'T' sequences converts to seasonal. Interestingly, 67.39 % of dataset is pandemic. As the result, these rules alone classify almost all sequences correctly. In other words, the generated rules can cover all pandemic part.

Table 3 shows potential mutations of M1 protein sequences which are unraveled by Ripper algorithm. Ratio of seasonal sequences for M1 protein was 30.73%, which is similar to support of all rules related to seasonal class (30%). It means, all seasonal sequences were covered by these rules. For example, the first rule of Table 3 expresses that in 29.43% of M1 sequences, when position 147 is I, the sequence is seasonal. In fact, when position 147 of M1 is I, the sequences are seasonal. Rules related to pandemic class can be interpreted similarly.

Also, supplementary tables present some rules which govern sequences of the other segments. For example, when position 334 of NP is H, almost certainly the sequence is pandemic. We can identify seasonal sequences when position 364 of PB1 is L. These rules were extracted by different algorithms and some of them are complementary to each other. It means one rule covers related class completely; therefore, the complement of that rule also covers the opposite class.

In addition, we defined a function based on extracted rule of each segment similar to following equation:

$$Pandemic(seq) = \begin{cases} 1 & for\ pandemic\ seq \\ 0 & for\ \ seasonal\ seq \end{cases}$$

If more than two protein sequences exist, and the following equation is satisfied, the result will be more reliable and accurate than when just one protein sequence exists.

$$\sum_{k=1}^{no.protein\ that\ seq\ exist} Pandemic(seq) > \frac{k}{2}$$

**Biological interpretation of identified rules (mutations) of different segments**

The biological importance of some of the discovered positions (rules) is presented at Table 4. For some of mutations, Figure 2 illustrates the contribution of each discovered mutation point (rule) in each segment in discriminating of pandemic sequences from seasonal ones, independently. Supplementary Tables confirm the Figure 2 where discrimination frequencies are above 90%. That means, these positions discriminated sequences with high accuracy.

Furthermore, as presented in Table 3 and Supplementary Tables S1-S9, we report several associative rules (combination of mutation positions). Noticeably, when two important mutation points join each other (associative rules), they cover the related part more comprehensively. So, associative rules help us detect pandemic sequences more accurately.

In order to gather pandemic markers in all segments, we select strains that sequences of all segment were available. Finally, 3723 sequences have remained which 1000 of them is seasonal and 2723 are pandemic. We put potential markers of each segments beside each other and select the markers with unique and significant role. Figure 3 shows these positions together. As it can be inferred from Figure 3, 30 markers of HA, NA, M1, PB1, PB2, PA, and NP segments can identify pandemic and seasonal strain accurately. The amino acid characteristics of these 30 markers are quite different in more than 92% sequences (frequency of row 1, 2 of Figure 3 were 95.51% and 92.60% for pandemic and seasonal sequences respectively). We discovered a pattern distinguishing seasonal strains from pandemic ones.

## Discussion

Pandemic mutational markers are not limited to the surface proteins (Hemagglutinin and Neuraminidase) and can be investigated in internal segments as well as combination of internal and external segments. In this study, for the first time, mutation points in all segments of H1N1 influenza viruses were detected in a large scale. The rules (hot spots) were extracted from more than 4000 sequences. Visual alignment was not able to statistically detect the association rule (co-occurrence) between mutations. The proposed machine learning based approach successfully addressed the shortcoming and discovered the co-occurrence of mutations in different segment.

For the first time, we determined potential hot spots by whole genome and proteome analysis in a large scale representing the discriminative power of of both external and internal mutations on pandemic discrimination. Influenza A evolves through different mechanisms, including point mutations and gene reassortment causing antigenic drift and antigenic shift respectively (Suzuki, 2005). Interactions occur between viruses of different lineages. The segmented structure of the virus facilitates gene reassortment when viruses from different hosts simultaneously infect a single cell (Ebrahimi et al., 2014b). The reassortment of genetic material between viruses with different host origins can significantly alter antigenic sites (Brockwell- Staats et al., 2009). By this mechanism, novel viruses may enter the human population that lacks previous immunity, potentially causing the emergence of pandemics or disastrous epidemics (CHENG, 2006). We highlight this point that the effect of whole segments in emergence of pandemic influenza needs to be considered.

Three global pandemics in the 20th century emerged by antigenic shift between viruses with different host origin. The 1957 H2N2 pandemic was the consequence of a reassortant of five human H1N1 segments and avian segments encoding the viral surface proteins and the PB1 protein. Similarly, the 1968 H3N2 pandemic involved a reassortment of avian segments encoding hemagglutinin and PB1 (Kilbourne, 2006). The viral genome of the 2009 H1N1 pandemic had a more complex reassortment history involving triple reassortment between hosts which mixed segments of human H3N2 (PB1), avian influenza A virus (PA, PB2) and classical North American swine influenza A virus (HA, NP, NS) (Garten et al., 2009; Smith et al., 2009). This genetic reassortment pattern allowed virus to infect human, swine, and birds and, in addition, it acquired the life-threatening ability to transmit from human to human without the need to intermediate swine or bird.

In previous studies, the importance of other segments (except HA, NA) was majorly ignored. Also, mutations were limited to a specific location and a few numbers of sequences were considered in previous studies. Here, we discovered the potential marker positions in a large scale study. We documented that other proteins of influenza virus can accurately identify pandemic phenotype even in absence of HA or NA segments. Our approach could discriminate sequences to pandemic and seasonal groups with more than 95% accuracy for PA and NP, 99.33% accuracy for NA and 100% accuracy for M1, M2, PB1, NS1, and NS2. If sequences of all segments of one strain are available, synchronously, the accuracy of our recognition will reach 100%.

Machine learning has offered new possibilities in virus research such as predicting the outcome of therapy based on viral nucleotide attributes (KayvanJoo et al., 2014) and unravelling the underlying layers of subtype differentiation (Ebrahimi et al., 2014b). For discriminative pattern discovery between pandemic and seasonal sequences, CBA algorithm outperformed the other machine learning models. The distinguished power of CBA algorithm to discover and combine the mutilations from different segments of influenza for distinguishing of pandemic sequences was remarkable in this study. In line with this finding, CBA has demonstrated high performance in identification of host range of influenza sequence (avian, human, and swine) by combination of mutation positions in all segments of influenza as host discriminative rules, leading to the establishment of a novel approach for identification of influenza virus host range and zoonotic transmissible sequences (Kargarfard et al., 2016). CBA is a high performance and robust classifier that integrates classification algorithm with association rule mining algorithm, the two key discriminative machine learning approaches techniques (Kargarfard et al., 2015). CBA find homogenous groups within heterogenous data, based on the minimum support. Then, CBA applies discriminative rules with high confidence in each homogenous group (Liu et al., 2001).We suggest to develop the similar model for H5N1, as well as mixture of all subtypes of influenza in future studies. Analysis of pre-pandemic strains (as a reference) in comparison with pandemic strains in future studies can contribute in increasing the power of discriminating rules.

**Conclusion**

Here, for the first time, we successfully applied rule based classification techniques to better distinguish between pandemic and seasonal influenza H1N1 based on whole segments of influenza. Rule based classification techniques provide the opportunity to first discover significant rules in respect of label

variable (pandemic and seasonal), and then to apply these rules in pandemic prediction of strains. Analysis of mutation positions in all segments of influenza genome as well as presenting a n accurate integrative pattern discovery algorithm (CBA model) discriminating pandemic from seasonal sequences by combination of mutated positions are the key point of the current study. The approach developed in this study can be employed in unraveling the underlying rules of influenza host range increase in future studies, as well as unraveling the underlying layers of pathogenicity in other viruses. The distinguished power of CBA algorithm to discover and combine the mutilations from different segments of influenza toward pandemic emergence is one of the highlights of this study, opening a new avenue for application of this advanced algorithm in biomedical research.

## Competing interests

The authors declare that there is no conflict of interest.

**Acknowledgements**

We are thankful from Dr. Morgan Newman and Professor Jeremy Timmis, School of Biological Sciences Science of The University of Adelaide, for English editing of the manuscript. The computational biology analyses of this study carried out at Shiraz University and The University of Adelaide.

## References


, 2015. RapidMiner.

Agrawal, R. and Srikant, R., 1994. Fast algorithms for mining association rules, Proc. 20th int. conf. very large data bases, VLDB. pp. 487-499.

Attaluri, P.K., Zheng, X., Chen, Z. and Lu, G., 2009. Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic. BIOT-2009, 21.

Bakhtiarizadeh, M.R., Moradi-Shahrbabak, M., Ebrahimi, M. and Ebrahimie, E., 2014. Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. Journal of theoretical biology 356, 213-222.

Baudin, F., Petit, I., Weissenhorn, W. and Ruigrok, R.W., 2001. In vitro dissection of the membrane and RNP binding activities of influenza virus M1 protein. Virology 281, 102-108.

Bing Liu, W.H., Yiming Ma, 1998. Integrating classification and association rule mining, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation). New York, USA.

Brockwell- Staats, C., Webster, R.G. and Webby, R.J., 2009. Diversity of influenza viruses in swine and the emergence of a novel human pandemic influenza A (H1N1). Influenza and other respiratory viruses 3, 207-213.

Chen, G.-W. and Shih, S.-R., 2009. Genomic signatures of influenza A pandemic (H1N1) 2009 virus. Emerg Infect Dis 15, 1897-1903.

CHENG, V.C., 2006. Human Swine Influenza. Education 11.

Daud, N.R. and Corne, D.W., 2009. Human readable rule induction in medical data mining, Proceedings of the European Computing Conference. Springer, pp. 787-798.

Du, Q.-S., Wang, S.-Q., Huang, R.-B. and Chou, K.-C., 2010. Computational 3D structures of drug-targeting proteins in the 2009-H1N1 influenza A virus. Chemical Physics Letters 485, 191-195.

Ebrahimi, M., Aghagolzadeh, P., Shamabadi, N., Tahmasebi, A., Alsharifi, M., Adelson, D.L., Hemmatzadeh, F. and Ebrahimie, E., 2014a. Understanding the Undelaying Mechanism of HA-Subtyping in the Level of Physic-Chemical Characteristics of Protein. PloS one 9, e96984.

Ebrahimi, M., Aghagolzadeh, P., Shamabadi, N., Tahmasebi, A., Alsharifi, M., Adelson, D.L., Hemmatzadeh, F. and Ebrahimie, E., 2014b. Understanding the underlying mechanism of HA-subtyping in the level of physic-chemical characteristics of protein. PloS one 9, e96984.

Ebrahimi, M., Ebrahimie, E., Shamabadi, N. and Ebrahimi, M., 2010. Are there any differences between features of proteins expressed in malignant and benign breast cancers? Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences 15, 299.

Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimie, E. and Ebrahimi, M., 2011. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. PloS one 6, e23146.

Ebrahimie, E., Ebrahimi, F., Ebrahimi, M., Tomlinson, S. and Petrovski, K.R., 2018. Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. Computers and Electronics in Agriculture 147, 6-11.

Ebrahimie, E., Nurollah, Z., Ebrahimi, M., Hemmatzadeh, F. and Ignjatovic, J., 2015. Unique ability of pandemic influenza to downregulate the genes involved in neuronal disorders. Molecular biology reports 42, 1377-1390.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32, 1792-1797.

Garten, R.J., Davis, C.T., Russell, C.A., Shu, B., Lindstrom, S., Balish, A., Sessions, W.M., Xu, X., Skepner, E. and Deyde, V., 2009. Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. science 325, 197-201.

Hadifar, F., Ignjatovic, J., Tarigan, S., Indriani, R., Ebrahimie, E., Hasan, N.H., McWhorter, A., Putland, S., Ownagh, A. and Hemmatzadeh, F., 2014. Multimeric Recombinant M2e Protein-Based ELISA: A Significant Improvement in Differentiating Avian Influenza Infected Chickens from Vaccinated Ones. PLOS ONE 9, e108420.

Hasan, N.H., Ebrahimie, E., Ignjatovic, J., Tarigan, S., Peaston, A. and Hemmatzadeh, F., 2016. Epitope Mapping of Avian Influenza M2e Protein: Different Species Recognise Various Epitopes. PLOS ONE 11, e0156418.

Hemmatzadeh, F., Sumarningsih, S., Tarigan, S., Indriani, R., Dharmayanti, N.L.P.I., Ebrahimie, E. and Igniatovic, J., 2013. Recombinant M2e Protein-Based ELISA: A Novel and Inexpensive Approach for Differentiating Avian Influenza Infected Chickens from Vaccinated Ones. PLOS ONE 8, e56801.

Horimoto, T. and Kawaoka, Y., 2005. Influenza: lessons from past pandemics, warnings from current incidents. Nature Reviews Microbiology 3, 591-600.

Hu, W., 2010. Novel host markers in the 2009 pandemic H1N1 influenza A virus. Journal of Biomedical Science and Engineering 3, 584.

Jamali, A.A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R. and Ebrahimie, E., 2016. DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug Discovery Today 21, 718-724.

Kargarfard, F., Sami, A. and Ebrahimie, E., 2015. Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm. Journal of biomedical informatics 57, 181-188.

Kargarfard, F., Sami, A., Mohammadi-Dehcheshmeh, M. and Ebrahimie, E., 2016. Novel approach for identification of influenza virus host range and zoonotic transmissible sequences by determination of host-related associative positions in viral genome segments. BMC genomics 17, 925.

Kass, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. Applied statistics, 119-127.

KayvanJoo, A.H., Ebrahimi, M. and Haqshenas, G., 2014. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC research notes 7, 565.

Kilbourne, E.D., 2006. Influenza pandemics of the 20th century. Emerging infectious diseases 12, 9.

Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A. and Lopez, R., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.

Li, W., Han, J. and Pei, J., 2001. CMAR: Accurate and efficient classification based on multiple class-association rules, Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, pp. 369-376.

Lin, D., Lan, J. and Zhang, Z., 2007. Structure and function of the NS1 protein of influenza A virus. Acta biochimica et biophysica Sinica 39, 155-162.

Liu, B., Ma, Y. and Wong, C.-K., 2001. Classification using association rules: weaknesses and enhancements, Data mining for scientific and engineering applications. Springer, pp. 591-605.

Liu, T. and Ye, Z., 2005. Attenuating mutations of the matrix gene of influenza A/WSN/33 virus. Journal of virology 79, 1918-1923.

Maurer-Stroh, S., Lee, R., Eisenhaber, F., Cui, L., Phuah, S.P. and Lin, R., 2009a. A new common mutation in the hemagglutinin of the 2009 (H1N1) influenza A virus. PLoS currents 2, RRN1162-RRN1162.

Maurer-Stroh, S., Ma, J., Lee, R.T., Sirota, F.L. and Eisenhaber, F., 2009b. Mapping the sequence mutations of the 2009 H1N1 influenza A virus neuraminidase relative to drug and antibody binding sites. Biology Direct 4, 18.

Miotto, O., Heiny, A., Albrecht, R., García-Sastre, A., Tan, T.W., August, J.T. and Brusic, V., 2010. Complete-proteome mapping of human influenza A adaptive mutations: implications for human transmissibility of zoonotic strains. PloS one 5, e9025.

Mohammadi-Dehcheshmeh, M., Niazi, A., Ebrahimi, M., Tahsili, M., Nurollah, Z., Khaksefid, R.E., Ebrahimi, M. and Ebrahimie, E., 2018. Unified Transcriptomic Signature of Arbuscular Mycorrhiza Colonization in Roots of Medicago truncatula by Integration of Machine Learning, Promoter Analysis, and Direct Merging Meta-Analysis. Frontiers in Plant Science 9.

Notredame, C., Higgins, D.G. and Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302, 205-217.

Ohtsu, Y., Honda, Y., Sakata, Y., Kato, H. and Toyoda, T., 2002. Fine mapping of the subunit binding sites of influenza virus RNA polymerase. Microbiology and immunology 46, 167-175.

Quinlan, J.R., 1993. C4. 5: programs for machine learning, Morgan kaufmann.

Rokach, L., 2008. Data mining with decision trees: theory and applications, World scientific.

Sharifi, S., Pakdel, A., Ebrahimi, M., Reecy, J.M., Fazeli Farsani, S. and Ebrahimie, E., 2018. Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. PLOS ONE 13, e0191227.

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M. and Ebrahimie, E., 2014. Determining the Most Important Physiological and Agronomic Traits Contributing to Maize Grain Yield through Machine Learning Algorithms: A New Avenue in Intelligent Agriculture. PloS one 9, e97288.

Sivathayalan, A., 2009. Comparison of Clustering and Classification Methods Combined with Dimension Reduction Using Gene Expression Data. Carleton University.

Smith, G.J., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J. and Bhatt, S., 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459, 1122-1125.

Squires, R.B., Noronha, J., Hunt, V., García-Sastre, A., Macken, C., Baumgarth, N., Suarez, D., Pickett, B.E., Zhang, Y., Larsen, C.N., Ramsey, A., Zhou, L., Zaremba, S., Kumar, S., Deitrich, J., Klem, E. and Scheuermann, R.H., 2012. Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. Influenza and Other Respiratory Viruses 6, 404-416.

Strengell, M., Ikonen, N., Ziegler, T. and Julkunen, I., 2011. Minor changes in the hemagglutinin of influenza A (H1N1) 2009 virus alter its antigenic properties. PLoS One 6, e25848.

Suzuki, Y., 2005. Sialobiology of influenza: molecular mechanism of host range variation of influenza viruses. Biological and Pharmaceutical Bulletin 28, 399-408.

Tarigan, S., Wibowo, M.H., Indriani, R., Sumarningsih, S., Artanto, S., Idris, S., Durr, P.A., Asmara, W., Ebrahimie, E., Stevenson, M.A. and Ignjatovic, J., 2018. Field effectiveness of highly pathogenic avian influenza H5N1 vaccination in commercial layers in Indonesia. PLOS ONE 13, e0190947.

Veljkovic, V., Niman, H.L., Glisic, S., Veljkovic, N., Perovic, V. and Muller, C.P., 2009. Identification of hemagglutinin structural domain and polymorphisms which may modulate swine H1N1 interactions with human receptor. BMC Structural Biology 9, 62.

William, C., 1995. Fast effective rule induction, Twelfth International Conference on Machine Learning. pp. 115-123.

Ye, Q., Krug, R.M. and Tao, Y.J., 2006. The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. Nature 444, 1078-1082.

Yin, X. and Han, J., 2003. CPAR: Classification based on Predictive Association Rules, SDM. SIAM, pp. 331-335.

Yuan, P., Bartlam, M., Lou, Z., Chen, S., Zhou, J., He, X., Lv, Z., Ge, R., Li, X. and Deng, T., 2009. Crystal structure of an avian influenza polymerase PAN reveals an endonuclease active site. Nature 458, 909-913.

## Figures

**Figure 1.** A schematic view of the proposed approach of knowledge discovery and prediction of H1N1 pandemic influenza by whole segments analysis and application of rule based classifier

**Figure 2.** The independent contribution of some of the discovered amino acid mutation positions in discrimination of 2009 H1N1 pandemic from seasonal ones. The figure includes 10 charts which any one of them represents important positions of each segment of influenza A sequences. Also, each chart represents what percentage of each position were varied in pandemic and seasonal sequences.

**Figure 3.** Significant mutated amino acid positions in all segments of influenza distinguishing 2009 seasonal H1N1 and 2009 pandemic H1N1. These amino acid mutation points are the combination of important markers of HA, NA, M1, PB1, PB2, and NP segments that are different in pandemic and seasonal sequences. The last column of figure reflects relative frequency of the combination in the sequenced genomes.

## Tables

**Table 1**. Number of external and internal sequence segments of human H1N1 influenza that were used in this study for rule discovery towards discrimination of pandemic influenza from seasonal ones.

| Protein name | No. Pandemic sequences | No .Seasonal sequences | No. Total sequences |
|---|---|---|---|
| HA | 3621 | 1752 | 5373 |
| NA | 3283 | 1633 | 4916 |
| PA | 3274 | 1145 | 4392 |
| NP | 3326 | 1147 | 4473 |
| PB1 | 3159 | 1146 | 4305 |
| PB2 | 3098 | 1128 | 4226 |
| M1 | 3417 | 1516 | 4933 |
| M2 | 3059 | 1490 | 4549 |
| NS1 | 3435 | 1119 | 4554 |
| NS2 | 3135 | 1105 | 4240 |

**Table 2.** Rules extracted from HA nucleotide sequences of human H1N1 strain discriminating pandemic from seasonal sequences and their and their confidence and support, using CBA (classification based on association rule mining).

| Class | Rule | Support | Confidence |
|---|---|---|---|
| Pandemic | Not (Att260 = 'T') | 67.39% | 100% |
| Seasonal | (Att260 = 'T') | 32.58% | 100% |

**Table 3**. Rules extracted from M1 protein of human H1N1 influenza in discrimination of pandemic sequences from seasonal ones and their confidence and support, using Ripper algorithm.

| Class | Rule | Support | Confidence |
|---|---|---|---|
| Seasonal | Att147 = 'I' | 29.43% | 100% |
| Seasonal | Att160 = 'K' | 29.33% | 100% |
| Seasonal | Att101 = 'R' | 30.67% | 99.73% |
| Seasonal | Att166 = 'V' | 30.63% | 99.66% |
| Seasonal | Att227 = 'T' | 29.83% | 98.09% |
| Pandemic | Att166 = 'A' and Att203 = 'M' | 68.84% | 99.76% |
| Pandemic | Att137 = 'T' | 69.53% | 99.50% |
| Pandemic | Att207 = 'N' | 70.50% | 98.13% |
| Pandemic | Att160 = 'R' | 70.64% | 98.04% |
| Pandemic | Att227 = 'A' | 70.13% | 97.94% |

**Table 4.** Biological importance of some of the mutation positions of this study in determination of pandemic influenza.

| Segment | Positions | Comment (biological importance) |
|---|---|---|
| HA | 274 , 286 | D274 are predicted to be "hot-spot" for polymorphisms which could increase infectivity of A/H1N1 virus. The domain 286-326 was identified to be involved in virus/receptor interaction (Veljkovic et al., 2009) |
| M1 | 101,137 ,207 | The functions of 101RKLKR105 were investigated by introducing substitution into the M gene of influenza virus A/WSN/33. Mutations, R101S or R105S, had effect on viral replication (Liu and Ye, 2005). Position 137 was detected as avian-human host shift markers (Miotto et al., 2010). Positions 207 and 209 were in the C-terminal part of M1 (residues 165-252) that binds to vRNP (Baudin et al., 2001). |
| M2 | 43,50 | Position 43 is the possible binding site (Thr43) for the inhibitors (adamantane-based Drugs (Du et al., 2010). Position 50 was avian-human host shift sites (Chen and Shih, 2009) |
| PB1 | 12, 211 618,728 | PB1 can binds to viral promoter and interact with PB2, NP, and PA. Position 12 within PB1-PA binding domain (residues 1-25) and two position 618 and 728 in the PB1-PB2 binding domain (residues 600-757) were reported (Ohtsu et al., 2002) (Hu, 2010). A mutation occurred at position 211 on H1N1 human influenza at New Zealand, Australia U.S.A., Asia (Daud and Corne, 2009) |
| NS1 | 111,112 | Positions 111,112 were in the effector domain. NS1 is a multifunctional protein contained in both protein-protein and protein-RNA interactions. Its C-terminal region (residues 74-237) contains the effector domain that prevent the substitution and exportation of the host cellular antiviral mRNAs (Lin et al., 2007). |
| NS2 | 57 | At (Hu, 2010) position 57 was reported as a high significant marker like swine-human host switch marker. |
| NA | 134, 174, 265, 296, 297 | The antigenic sites of N1 are residues 83-143, 156-190, 252-303, 330, 332, 340-345, 368, 370,387-395,431-435,448-468. So sites 134, 174, 265, 296, and 297 are were at the antigenic sites of N1 (Maurer-Stroh et al., 2009b). |
| PA | 257,363 | Two position 257 and 363 were in the C-terminal domain of PA (residues 257-716) which binds to PB1 for complex formation and nuclear transport (Yuan et al., 2009). |
| NP | 334 | The overall structure of nucleoprotein divided into two domains: a head and a body. The body domain of NP includes the binding positions for the viral polymerase. It is formed by residues 21–149, 273–396 and 453–489. So position 334 is a binding site (Ye et al., 2006). |

## List of abbreviations

CBA: *Classification Based on Associations*

CMAR: *Classification based on Multiple Association Rules*

CPAR: *Classification based on Predictive Association Rules*

**Highlights**

- Knowledge extraction in influenza pandemic strains based on all segments
- Rule based classification for discovery of mutation markers of pandemic influenza
- Pattern discovery for discriminating pandemic strains from seasonal ones by machine learning

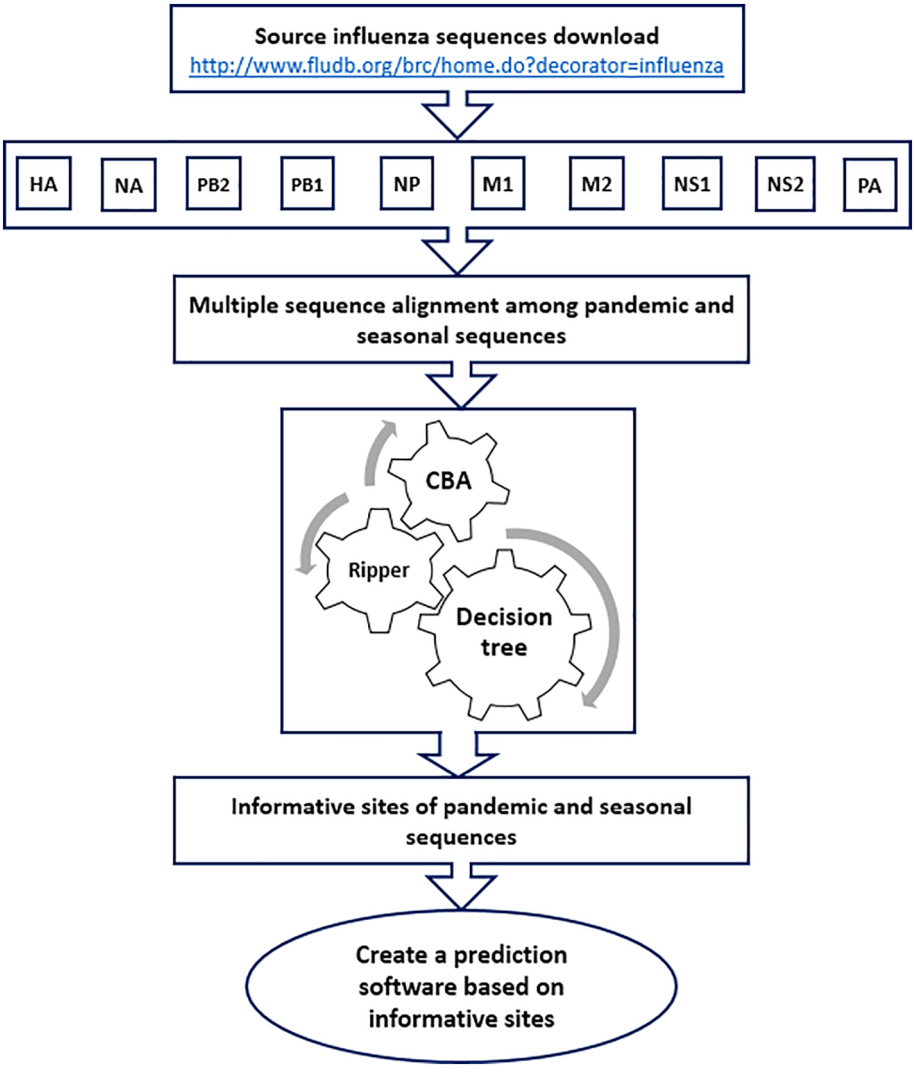
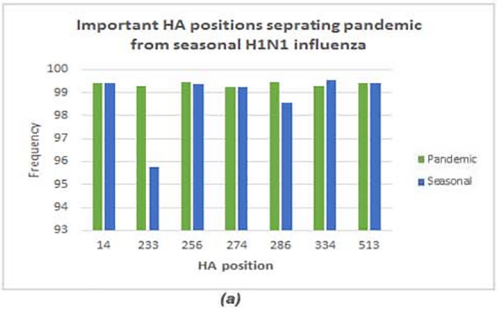
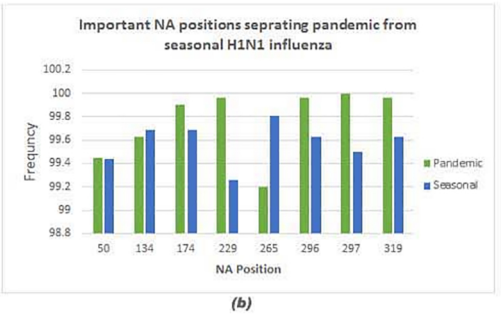

Source influenza sequences download
http://www.fludb.org/brc/home.do?decorator=influenza
HA
NA
PB2
PB1
NP
M1
M2
NS1
NS2
PA
Multiple sequence alignment among pandemic and seasonal sequences
CBA
Ripper
Decision tree
Informative sites of pandemic and seasonal sequences
Create a prediction software based on informative sites

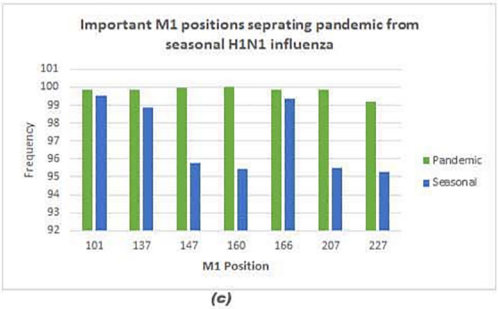
Figure 1



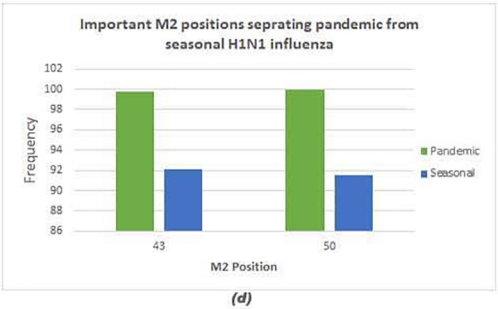

(a)



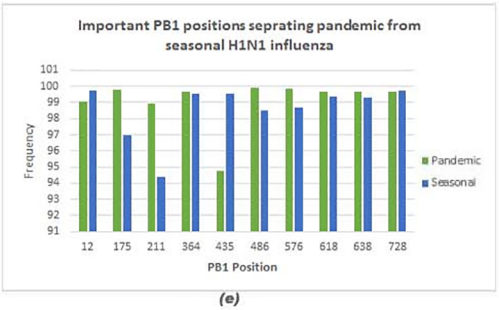

(b)



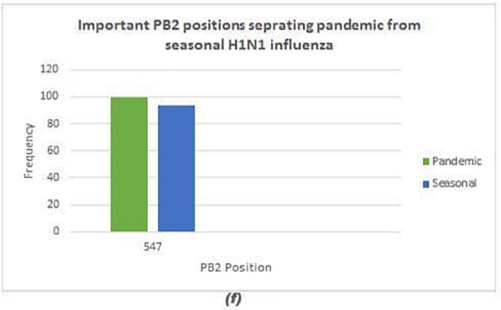

(c)



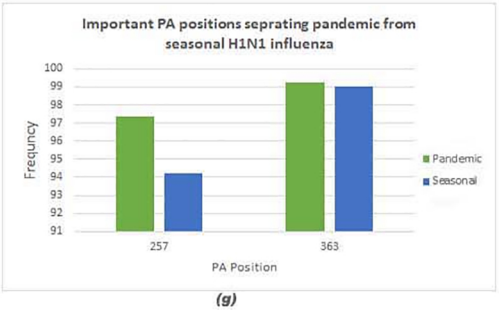

(d)



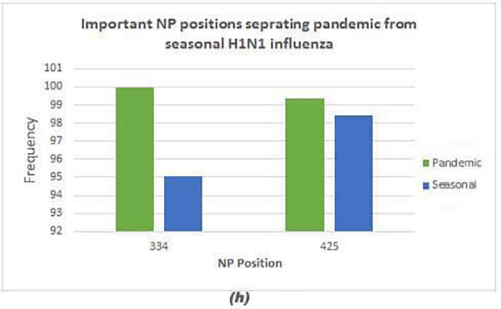

(e)



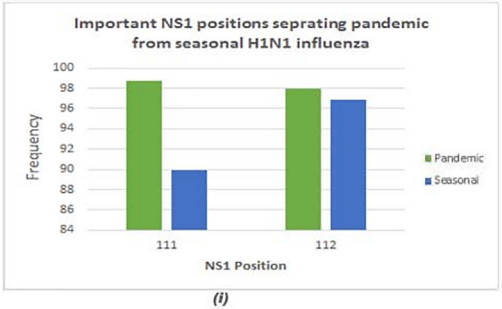

(f)



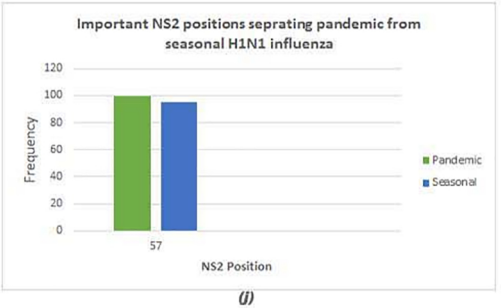

(g)




(h)




(i)




(j)

Figure 2

| HA_233 | HA_256 | HA_274 | HA_286 | HA_334 | HA_513 | NA_50 | NA_134 | NA_174 | NA_229 | NA_265 | NA_296 | NA_297 | NA_319 | M1_101 | M1_137 | M1_147 | M1_160 | M1_166 | PB1_12 | PB1_175 | PB1_486 | PB1_576 | PB1_618 | PB1_638 | PB1_728 | PB2_547 | PA_363 | NP_334 | NP_425 | TYPE | Frequency |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | K | M | D | L | A | I | P | V | N | R | I | T | E | K | T | V | R | A | I | N | K | L | D | D | V | V | R | H | V | Pandemic | 95.51% |
| K | T | L | N | M | S | T | H | A | K | K | V | M | D | R | A | I | K | V | V | D | R | I | E | E | I | I | K | N | I | Seasonal | 92.60% |
| I | K | M | D | L | A | I | P | V | N | R | I | T | E | K | T | V | R | A | I | N | K | L | D | D | V | V | R | H | I | Pandemic | 0.70% |
| I | K | M | D | L | A | I | P | V | N | K | I | T | E | K | T | V | R | A | I | N | K | L | D | D | V | V | R | H | V | Pandemic | 0.60% |
| I | K | M | D | L | A | I | P | V | N | R | I | T | E | K | T | V | R | A | I | N | K | L | D | D | V | V | K | H | V | Pandemic | 0.55% |
| I | K | M | D | L | A | T | P | V | N | R | I | T | E | K | T | V | R | A | I | N | K | L | D | D | V | V | R | H | V | Pandemic | 0.40% |
| K | T | L | N | M | S | T | H | A | K | K | V | M | D | R | A | V | R | V | V | N | R | I | E | E | I | V | K | H | I | Seasonal | 1.10% |
| I | K | M | D | L | A | I | P | V | N | R | I | T | E | K | T | V | R | A | I | N | K | L | D | D | V | I | R | H | V | Pandemic | 0.30% |
| V | K | M | D | L | A | I | P | V | N | R | I | T | E | K | T | V | R | A | I | N | K | L | D | D | V | V | R | H | V | Pandemic | 0.30% |
| K | T | L | N | M | S | T | H | A | K | K | V | M | D | R | A | I | K | V | V | D | R | I | E | E | I | V | K | N | I | Seasonal | 0.80% |
| E | T | L | N | M | S | T | H | A | K | K | V | M | D | R | A | I | K | V | V | N | R | I | E | E | I | I | K | N | I | Seasonal | 0.80% |

Figure 3