Identification of potential critical virulent sites based on hemagglutinin of influenza A virus in past pandemic strains

Yin, Rui; Fransiskus, Xaverius Ivan; Zheng, Jie; Zhou, Xinrui; Chow, Vincent T. K.; Kwoh, Chee Keong

2017

Yin, R., Zhou, X., Fransiskus, X. I., Zheng, J., Chow, V. T. K., & Kwoh, C. K. (2017).
Identification of potential critical virulent sites based on hemagglutinin of influenza A virus
in past pandemic strains. Proceedings of the 6th International Conference on
Bioinformatics and Biomedical Science - ICBBS '17. doi:10.1145/3121138.3121166

<https://hdl.handle.net/10356/102511>

<https://doi.org/10.1145/3121138.3121166>

© 2017 Association for Computing Machinery (ACM). All rights reserved. This paper was
published in Proceedings of the 6th International Conference on Bioinformatics and
Biomedical Science and is made available with permission of Association for Computing
Machinery (ACM).

Downloaded on 26 Apr 2023 22:21:28 SGT

Identification of potential critical virulent sites based on hemagglutinin of influenza A virus in past pandemic strains

Rui Yin ¹
YINR0002@e.ntu.edu.sg

Xinrui Zhou ¹
XZHOU010@e.ntu.edu.sg

Fransiskus Xaverius Ivan ¹
FIVAN@ntu.edu.sg

Jie Zheng ^{1, 2}
ZhengJie@ntu.edu.sg

Vincent T. K. Chow ³
micctk@nus.edu.sg

Chee Keong Kwoh ^{1,*}
ASCKKWOH@ntu.edu.sg

¹ School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798 Singapore, Singapore.

² Genome Institute of Singapore, A*STAR, Biopolis, 138672 Singapore, Singapore.

³ Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, 5 Science Drive 2, MD4, Level 3, Kent Ridge, Singapore 117545

* Correspondence: ASCKKWOH@ntu.edu.sg

ABSTRACT

The influenza pandemics have caused millions of deaths and enormous economic loss. Current circulating influenza viruses in human, avian, swine and other animals are potential to evolve into novel strains that may cause another pandemic in the future. Hence, recognizing the determinants of pandemic strains helps to raise the alarm of future pandemics. With increasingly huge biological data, computational modeling is a good technique for analyzing data, providing novel insight into significant patterns and rules. Here we define a binary classification problem of categorizing influenza strains into pandemic and non-pandemic classes based on amino acid sequences. Three rule-based algorithms are applied, namely OneR, JRip and PART, to extract rules, composed of potential critical virulent sites. The results present good performance in term of accuracy, specificity, sensitivity and F-measure (more than 0.9 on average for each). Fourteen out of the sixteen potential critical virulent sites detected in our experiments are overlapped with receptor binding sites or antigenic sites. In addition, some variations occurred in these sites are known to affect the pathogenicity of influenza viruses or to cause more severe symptom in the infected patients. The pandemic potential of uncovered sites in our study needs to be further experimentally validated.

Keywords

Influenza A virus; Pandemic; Hemagglutinin; Potential virulent sites; Rule-based classification.

1. INTRODUCTION

The influenza A virus is a member of the Orthomyxoviridae family, containing a genome makeup of eight single-stranded, negative-sense RNA segments. It may cause fever, headaches, sore throat, body aches, nose congestion, bronchitis, pneumonia and even death when humans get infected [1]. Influenza infection becomes an epidemic when it rapidly spreads to many people within a short period of time in a given population, which occurs

every year around the world [2]. While an epidemic is usually restricted to the locations, if it spreads to other countries or continents and affects numerous people, then it will turn to a pandemic. There are mainly five influenza pandemics since the nineteenth century including 1918 H1N1 pandemic in Spain, 1957 H2N2 Asian flu, 1968 H3N2 Hong Kong flu, 1977 H1N1 Russian flu and 2009 swine-origin H1N1 China flu, which caused millions of people's deaths as well as enormous economic loss, intensively threatening people's health and safety. In 2009, WHO made restatement description of pandemic to make clear the meaning of the word "pandemic" and the way to recognize it, i.e. "An influenza pandemic may occur when a new influenza virus appears against which the human population has no immunity [3]." Recently, a new influenza type H7N9 appeared in China, cumulative reported human infections with 359 deaths out of 918 cases by 16 Jan, 2017. People are concerned about its potential of triggering another pandemic of this virus and the risk of impacting public health. Therefore, the identification of critical sites from previous pandemic strains that caused human diseases with high virulent influenza is a very important step in promoting the surveillance system and protecting public safety from possible future pandemic.

The determinants of pandemic influenza strains remain an open and challenging problem. Animal models, such as mice, ferrets and non-human primates, have been significant helpers in analyzing the host-range, pathogenicity and virulence (from mild to severe) of influenza virus. Guinea pigs are also often used for transmission studies. Maines et al. analyzed the transmissible ability of swine-origin 2009 A/H1N1 by establishing ferrets and mice model, shedding light on the atypical symptoms, including gastrointestinal distress and vomiting [4]. Some work demonstrated the increased virulence or enhanced transmission with acquisition of specific mutations. For example, Tokiko et al. showed an increased pathogenicity in macaques when conferred the avian-type receptor binding ability by HA-222D and HA-222G [5]. Sander et al. selected mutant A/H5N1 viruses,

indicating Q222L and G224S changed the receptor binding of H2 and H3 avian influenza binding specificity to alpha (2,6) linked sialic acid, which contributed to the outbreak of 1957 and 1968 pandemics [6]. Chinh et al. found Mutation R289K-induced conformational in H7N9 suggests potential adaptations of the virus itself for future drug-resistance [7]. Although experiments have detected several determinants of pandemic influenza strains, these animal models are expensive, difficult to work with and it is hard to analyze all factors thoroughly.

On the contrary, computational techniques enable researchers to discover knowledge in a more efficient way. Data mining approaches has been demonstrated to be more powerful than manpower in some fields including biology. Large-scale sequences available makes it possible to retrieve clues from past strains. Miotto et al. identified a catalogue of sites as human-to-human transmission markers [8], and Chen et al. explored the mutation co-occurrence in HA1 sequences and detected co-mutation sites under strong selective pressure, predicting the potential drifts with specific mutations of the viruses [9].

However, those works do not relate to the virulence of influenza directly. Here, we tried to explore the critical sites of virulent strains by defining it as a binary classification problem. We labeled sequences from recent pandemics as high virulent, while the seasonal flu or epidemics as mild or less virulent. By applying three typical rule-learning based classification algorithms, namely OneR, JRip and PART, we classified the dataset with high accuracy. In addition, rules generated by those algorithms can shed light on the crucial virulent sites and help their detection. To evaluate the results, we turned to experimental results with animal models and found that 14 out of 16 detected sites are either located in epitope regions or relevant to the binding site to the host. Therefore, the rest are also indicated to be highly potential to play an important role in the virulence of an influenza strain.

2. Materials and Methods

2.1 Overview

We first collected hemagglutinin (HA) sequences of past pandemic and non-pandemic strains with three subtypes of influenza, namely, H1N1, H2N2 and H3N2. Then the data cleaning and preprocessing was carried out to acquire qualified strains. Next, three benchmark rule-based algorithms were applied to classify pandemic strains from non-pandemic strains using H1N1 as training and testing dataset by ten-fold cross validation. Some rules and potential critical sites related to the virulent pandemic strains were obtained. Finally, these sites were mapped into H2N2 and H3N2 PDM strains based on amino acid distribution. The workflow of the procedures is summarized in Figure 1.

2.2 Data collection

Influenza sequences were obtained from Influenza virus Resource(IVR;<https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database>) on 31 Dec, 2016. These amino acid sequences were retrieved of HA segment with full length plus and human host from 1918 to 2016. The experimental strains can be categorized into two types, pandemic and non-pandemic classes. The pandemic class, denoted as “PDM”, includes sequences from five pandemics in recent centuries, where subtypes

are set as the corresponding pandemic virus subtype, namely H1N1 for 1918 Spanish flu, 1977 Russian flu and 2009 swine originated influenza virus, H2N2 for the 1957 Asian flu and H3N2 for 1968 Hong Kong pandemic. Viruses could circulate for a period of time even after the pandemic although with much less incidence. We labeled sequences from two consecutive years of each pandemic as PDM sequence. Correspondingly, the dataset collected for the non-pandemic class, denoted as “NONPDM” was made up of all the strains of subtype H1N1, H2N2 and H3N2 from 1918 to 2016 except for the pandemic strains, during which there was only some small-scale outbreaks that happened seasonally. The datasets ended up with 3275 samples (collapsing identical sequences) for PDM class, with 3216 H1N1 strains, 35 H2N2 strains and 24 H3N2 strains respectively. Meanwhile, we collected 8989 samples for the NONPDM class, with 3329 H1N1 strains, 45 H2N2 strains and 5615 H3N2 strains. The details of datasets collected for each class are in Table 1.

Fig 1. The flowchart for binary classification of pandemic and non-pandemic strains.

Table 1. Data collection of subtype H1N1, H2N2 and H3N2 in pandemic and non-pandemic years of recent centuries.

Period	Name	Circulating	Subtype	Number	Class
1918-1919	Spanish flu	1950s	H1N1	1	PDM
1957-1958	Asian	1960s	H2N2	35	
1968-1969	Hong Kong	circulating	H3N2	25	
1977-1978	Russian	circulating	H1N1	14	
2009-2010	China	circulating	H1N1	3201	NON PDM
1920-1976	--	--	H1N1	54	
1979-2008	--	--	H1N1	1072	
2011-2016	--	--	H1N1	2203	
1918-1956	--	--	H2N2	0	
1959-2016	--	--	H2N2	45	
1918-1967	--	--	H3N2	0	
1970-2016	--	--	H3N2	5615	

2.3 Data cleaning and preprocessing

Due to the different length of each subtype of HA, we have classified strains from multiple clades of H1, H2 and H3. We first assembled PDM and NONPDM strains of H1N1. Multiple sequence alignment was implemented by MAFFT (Multiple Alignment using Fast Fourier Transform) with FASTA output format in the same order as input samples. FFT-NS-2 mode was selected as running strategy. The results showed many deletions and insertions of aligned sequences because of different HA length in H1N1, which turned out to have great impact in the distribution and alignment of amino acid sites. As the length of 566 amino acids (aa) is the most common length for a complete HA segment of H1N1 of influenza A virus, we use this length as a criterion to eliminate the strains that have more than 566 aa. (The strains removed from the raw dataset are in the supporting materials File S1.) The remaining samples of H1N1 include 3201 PDM strains and 3327 NONPDM strains. We repeated multiple sequence alignment again by MAFFT with the same parameter settings. This process was also applied to subtype H2N2 and H3N2 by using MAFFT with one H3N2 PDM strain removed.

Comparison of residues between subtypes of influenza virus has been increasingly used for comparative studies across subtypes. An analysis of N-terminal cleavage sites for thirteen subtypes of influenza A hemagglutinin sequences has been described by Nobusawa and colleagues [10]. F.Burke and J.Smith extended this work for eighteen subtypes [11]. The analysis of known structures of HA of influenza virus allows us define structurally and functionally equivalent amino acids across all subtypes using a numbering system based on mature HA sequences. The N-terminal signal peptide cleavage site of HA was predicted using *signalP* for all HA subtypes [12]. On account of this numbering system, we were able to unify equivalent sites across subtypes H1, H2 and H3 of influenza A by deleting signal peptides. The three representatives with N-terminal sequences of mature HA proteins starting with “DTICIGYHANNSS”, “DQICIGYHANNSS”, and “QDLPGNDNSTATLCLGH-HAVPN” are “A/California/04/2009/H1N1pdm”, “A/Singapore/1/1957/H2N2” and “A/AICHI/2/68/H3N2” respectively [11]. The preprocessed datasets of aligned sequences including H1N1, H2N2 and H3N2 will be provided as File S2 in supporting materials.

For each sequence in the preprocessed dataset, we extracted every site as one feature, concatenated with the amino acid before its position (for example T190). If there was an insertion or deletion, “-” would be used for replacement. As a result, there are 549, 547 and 550 features for every sequence of H1N1, H2N2 and H3N2 subtypes respectively, followed by the classification type at the end of each sequence. The feature extraction for preprocessed sequences of these three subtypes are available as File S3 in supporting materials.

2.4 Binary classification

Machine learning has been widely applied in bioinformatics for many years and the diverse range of rapidly expanding data produced by modern molecular biology has fueled a need for accurate classification and prediction algorithms [13]. Aiming at our binary

classification issue that detecting critical virulent sites in PDM strains, rule-based machine learning method was selected for identifying a set of context-dependent rules that collectively store and apply knowledge in a piecewise manner to make predictions [14]. However, there is no unique classifier or rule that can be used to outperform others for all problems. Here, we applied three standard rule-based classification algorithms, namely, OneR, JRip and PART, to integrate the results of all the classifiers to complement each other. Comparative analysis on their performance of different classifiers will also be described. These methods have been implemented in Weka [15].

OneR, short for “One Rule”, is a simple and accurate learning classification algorithm that generates one rule for each predictor in the data. It only selects one rule with the smallest total error. The drawback of this method is that it can only identify a single feature that matters most in the dataset without informing more rules and sites that are potentially related. The second algorithm JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is based on association rules with reduced error pruning, a very common and effective technique in decision tree algorithms. Rules are optimized by generating and pruning two variants of each rule from randomized data, from which the one with the minimal decision length is selected. Finally, rules that would increase the decision length were deleted from the ruleset and the optimal ruleset is obtained. PART (Projective Adaptive Resonance Theory) is the third algorithm that uses partial decision trees to generate the decision combining two dominant rule mining strategies C4.5 and RIPPER, which is straightforward with no need for global optimization by either discarding or adjusting individual rules.

Considering different lengths of H1N1, H2N2 and H3N2 features as well as sample distributions for the three subtypes, H1N1 samples manifested best quality with sufficient numbers of PDM and NONPDM strains in a balanced way, compared with the other two subtypes, for which there is a lack of enough PDM strains or imbalanced class distribution. Therefore, we used H1N1 samples as input dataset. Ten-fold cross validation was applied to identify the rules and sites that are related to pandemic strains in H1N1 subtype. The detected sites were mapped across H2 and H3 subtypes with equivalent sites by a numbering system based on mature HA sequences. Overlapped sites were to be found using H2N2 and H3N2 strains based on amino acid distribution.

3. Results

The results show that these three classification algorithms are competitive with high performance. The performance metrics (i.e., including precision, recall, F-measure and ROC area) were calculated following the testing of the classifiers. To measure the proportions of positives and negatives that are correctly identified, we also calculated sensitivity and specificity to better understand the performance in the context of PDM and NONPDM classification. Sequences in the PDM class were taken as positive instances, while the NONPDM sequences were negative instances. Table 2 shows the details of performance for the classifiers.

Table 2. Performance of OneR, JRip and PART

	OneR	JRip	PART
Precision	0.883	0.926	0.921
Recall	0.881	0.926	0.921
Sensitivity	0.910	0.928	0.923
Specificity	0.858	0.925	0.919
F-measure	0.881	0.926	0.921
ROC area	0.880	0.952	0.971

The rules were discovered by OneR, JRip and PART with H1N1 datasets. OneR selected only one feature, site 185

that gets the smallest total error with 0.883 precision. JRip discovered six rules for the classification of PDM strains with twenty different sites identified with 0.926 precision, which performed the best among these three methods. In the method of PART, which is based on partial decision trees, generating more rules than the previous two methods. Here we only presented top two rules classified as PDM by PART that occupied most of the PDM strains with the precision of 0.921, ignoring NONPDM rules that were less important with the identification of critical virulent sites. Details of rules and detected sites of these three methods are shown in Table 3.

Table 3. Rules and sites generated by OneR, JRip and PART

	Rules (x/y) ^a	Sites
OneR	G185 -> PDM, N185 -> PDM, S185 -> PDM, V185 -> PDM A185 -> NONPDM, D185 -> NONPDM, I185 -> NONPDM, K185 ->NONPDM, M185 -> NONPDM, P185 -> NONPDM, R185 -> NONPDM, T185 -> NONPDM (5751/775)	185
JRip	S185 & E374 & A186 & S183 -> PDM (1846.0/17.0) S185 & I216 & V272 & K163 & S451 & A134 -> PDM (730.0/23.0) T190 & K274 -> PDM (441.0/169.0) E499 & S143 & D97 & E374 & S162 -> PDM (91.0/17.0) E499 & S143 & D97 & V520 & S162 & P271 -> PDM (192.0/66.0) S451 & T190 & K274 & D187 & V153 -> PDM (20/0) ~ ^b -> NONPDM (3209.0/180.0)	97, 134, 143, 153, 162, 163, 183, 185, 186, 187, 190, 216, 271, 272, 274, 374, 451, 499, 520
PART ^c	S185 & K239 & I216 & T133 & A521 & N473 & I372 & K494 & T281 & T25 & K146 & A423 & L526 & T13 & A134 & S143 & V272 & S451 & S263 & N228 & K169 & S71 & E491 & K154 & K163 & K308 & G202 & V234 & H273 & Q223 -> PDM (2195.0/11.0) E374 & K458 & V479 & N31 & D97 & K409 & N540 & I266 & R259 & K302 & I510 & S190 & A197 & S84 -> PDM (339.0/12.0)	13, 25, 31, 71, 84, 97, 133, 134, 143, 146, 154, 163, 169, 185, 190, 197, 202, 216, 223, 228, 234, 239, 259, 263, 266, 272, 273, 281, 302, 308, 372, 374, 409, 423, 451, 458, 473, 479, 491, 494, 510, 521, 526, 540

^a x stands for the number of instances covered by the rule, while y is the number of misclassified instances.

^b ~ means all the other features.

^c We only present two main rules that identifying PDM class, for which the sites are important related to virulence.

On obtaining the sites from generated rules by using the HA sequences of H1N1 PDM and NONPDM datasets, we transferred these identified sites into equivalent ones that located in H2N2 and H3N2 by a numbering system based on mature HA sequence. Due to the limited strains in H2N2 (45 NONPDM strains and 35 PDM strains) and imbalanced strains in H3N2 (5614 NONPDM strains and 24 PDM strains), directly applying these rule-based algorithms in H2N2 and H3N2 datasets did not present many rules associated with critical sites, neither making consistent results with the identified sites in H1N1 experiments. Considering that, we mapped the sites identified in H1N1 into equivalent ones in H2N2 and H3N2 strains using cross-subtype numbering scheme proposed by Burke and Smith [12]. As a result, we were able to select the sites that also possibly influence the virulence of influenza strains using H2N2 and H3N2 datasets based on following two constraint conditions. The first constraint is that the type of amino acid for each site in H2/H3 PDM strains is almost the same, but different from most amino acids in NONPDM strains in the same site. The second constraint is the amino acid existing in the PDM strains of each site is not the majority or the second majority amino acid. After cross-type site transformation and selection, we can get some critical sites generated by rule-based classification that not only existing in H1N1 subtype, but also make sense in H2N2

or H3N2 subtypes. The selection sites of different subtypes are presented in Table 4.

Table 4. Summary of features in the generated rules across H1N1, H2N2 and H3N2 of PDM strains (H1 numbering system)

	Features (a)*	b/a
OneR	185	1/1
JRip	134, 143, 153, 185, 186, 187, 273, 274, <u>451</u>	8/9
PART	134, 143, 185, 190, 202, 216, 223, 239, 259, <u>273, 451</u>, 540	10/12

* a stands for the number of features included in the extracted rules.

b is the number of features located in epitope regions. The features located on the epitope regions are in **bold** font. Features with reported effects but not on the epitope are underlined.

4. Discussion

Since its first emergence of pandemic after the nineteenth-century, the influenza A(H1N1) strain has been circulating and evolving up to now, during which several other pandemics occurred with new influenza subtypes, causing millions of human deaths and enormous economic loss, seriously affecting the public health. This study aims at

identifying some critical sites that are related to the pathogenicity of PDM strains.

The analysis performed in this study using HA gene of PDM and NONPDM sequences revealed the presence of sixteen distinct sites that may affect the pathogenicity, leading to the outbreak of pandemic. To evaluate the sites identified by the algorithms, we mapped all the sites on the HA presented in Table 4 and found 14 out of 16 are in epitope regions [16]. The results visualized in Figure 2 also show that the sites 134, 153, 186 and 187 identified by JRip and the sites 134 and 216 identified by PART are the receptor-binding sites among the 13 serotype HAs. The receptor specificity of HA has been considered as one of the determinants of the tissue tropism and host range in influenza virus [17]. Specific mutations on sites 134, 187, 190 and 223 could lead to increased virus binding to alpha (2,6) linkage [11]. In addition, some more sites are in four major antigenic regions defined by Brownlee and Fodor [18]: sites 185, 186 and 187 in Ca1 and site 202 in Sb. The accumulation of HA mutations by antigenic drift arising from population immune pressure is a significant cause of the emergence of new seasonal human influenza viruses [19].

Fig 2. Mapping the detected sites onto hemagglutinin of H1N1 virus (PDB ID: 4EDB [20]). The blue, red, cyan, yellow and green regions represent the five epitope regions A-E respectively. The spheres with corresponding numbers are our detected sites locating at the epitope regions.

Besides the sites that are receptor binding sites or antigenic sites, there are several residue positions and mutations that could be indicated by other experimental results. For example, the amino acid change S143G, which occupied 29/55 (52.7%) of viruses analyzed during 2009-2010 season, is located next to the antigenic site Ca [21]. Meanwhile, a previous report indicates that two viruses with the S143G substitution were observed to have reduced antigenicity against the A/California/07/2009 vaccine virus in the HI test [22]. Some amino acid changes in H3N2, H1N1 and H1N2 viruses between seasons in site 239 suggest that this site could have been through cross-type transmission. Mutation E274K occurring between 2004 and 2007 were maintained in the oseltamivir-resistant viruses that caused epidemics in the 2007-2009 seasons [23], which is also located next to the antigenic site 273. Substitution S451N, found by Antonio Piralla's et al., may increase the severity of symptom of infected

patients [21]. Annotations and references on those sites are summarized in Table 5.

Table 5. Summary of annotations of features bearing specific functions (H1 numbering system)

Features	Annotations	References
134	Receptor binding site, mutation S134A increased virus binding to alpha2-6 glycans	[11]
143	S143G observed reduced antigenicity against the A/California/07/2009 vaccine virus in the HI test	[22]
153	Receptor binding site	[11]
185	Antigenic site	[18]
186	Antigenic site, receptor binding site	[11], [18]
187	Antigenic site, receptor binding site, mutation E187G increased virus binding to alpha2-6 glycans	[11], [18]
190	Receptor binding site, mutation K190R increased virus binding to alpha2-6 glycans	[11]
202	Antigenic site	[18]
216	Receptor binding site	[11], [21]
223	Receptor binding site, mutation Q223L increased virus binding to alpha2-6 glycans	[11]
239	Amino acid changes in H3N2, H1N1, H1N2 viruses between seasons	[19]
273	Antigenic site	[18]
274	Mutation E274K were maintained in the oseltamivir-resistant viruses that caused epidemics during 2007-2009	[23]
451	Substitution S451N may increase the severity of infected patients	[21]

To sum up, three rule-based classification algorithms give good performance in classifying pandemic strains from non-pandemic. By retrieving experimentally determined functional features and epitope regions, we found that the detected sites are highly overlapped with the experimentally determined functional sites, mostly affecting the antigenicity or receptor binding mechanisms with host cells. When it comes to the virulence, the adaption of an influenza A virus to recognize host cells and immune responses are always considered critical measurements. Here, we focused on the virulence to human, and the binding mechanism to human cells will affect the pathogenicity to individuals and transmission efficiency among populations, both of which are key requirements of a pandemic. Avian influenza viruses prefer an alpha (2,3) sialic linkage, which are preferentially expressed on cells deeper in human lungs besides the avian intestinal epithelium. Normally an avian influenza virus has little chance to infect human, whose upper respiratory cells are mostly alpha (2,6) sialic

linkages, but it will cause more acute symptoms once humans get infected. However, when an avian influenza was conferred the ability to bind the alpha (2,6) linked sialic acid, it would be more pathogenic and contagious. Another factor, epitope, the antigenic determinant of viruses, is the main target recognized by the immune system. Thus, it makes significant sense that sites on epitope regions are classifiers between the pandemic and the non-pandemic strains.

In our findings, a few other sites detected, 451 and 540, are found outside the epitope regions or link to the binding of human host. No significant mutations are identified at site 540 affecting pathogenicity, which may indicate it is not as important as the top ones. However, it should also be noted that they are still potential key sites in causing pandemics. More biological experiments could be carried out on these sites with the guidance of computational techniques for further validation.

5. Conclusion

In this study, three rule-based machine learning algorithms have been used to identify potential virulent sites in the past pandemic strains, which prove to be a powerful and efficient way of biological knowledge discovery. Our results show a good performance with over 0.9 precision rate on average, detecting 16 potential virulent sites in total. The further validation confirmed that 14 out of 16 sites are located in epitope regions or relevant to the binding of host cells, which contributes a better understanding for rapid detection of genetic variants with potential of causing pandemics or epidemics. Nevertheless, some more sites identified in this study and the functions of specific mutations that are related to virulence, leading severe pandemic events, need to be further clarified.

6. Support materials

File S1 The list of accession number removed from the raw dataset of PDM and NONPDM strains. (DOCX)

File S2 The preprocessed datasets of aligned PDM and NONPDM sequences of H1N1, H2N2 and H3N2. (FAS)

File S3 Feature extracted from H1N1, H2N2 and H3N2 PDM and NONPDM strains. (CSV)

7. Acknowledgements

R.Y. conceived, performed experiments, interpreted results, and wrote the manuscript. X.Z, Ivan, J.Z., T.K.Chow and C.K. revised the paper, provided overall supervision, direction and leadership to the research.

8. Funding

This project is supported by AcRF Tier 2 grant MOE2014-T2-2-023, Ministry of Education, Singapore.

Conflict of Interest: none declared.

9. REFERENCES

[1] Monto A S, Gravenstein S, Elliott M, et al. Clinical signs and symptoms predicting influenza infection[J]. Archives of internal medicine, 2000, 160(21): 3243-3247.

[2] Boivin G, Hardy I, Tellier G, et al. Predicting influenza infections during epidemics with use of a clinical case definition[J]. Clinical infectious diseases, 2000, 31(5): 1166-1169.

[3] Doshi P. The elusive definition of pandemic influenza[J]. Bulletin of the World Health Organization, 2011, 89(7): 532-538.

[4] Maines T R, Jayaraman A, Belser J A, et al. Transmission and pathogenesis of swine-origin 2009 A (H1N1) influenza viruses in ferrets and mice[J]. Science, 2009, 325(5939): 484-487.

[5] Watanabe T, Shinya K, Watanabe S, et al. Avian-type receptor-binding ability can increase influenza virus pathogenicity in macaques[J]. Journal of virology, 2011: JVI. 00859-11.

[6] Herfst S, Schrauwen E J A, Linster M, et al. Airborne transmission of influenza A/H5N1 virus between ferrets[J]. science, 2012, 336(6088): 1534-1541.

[7] Su C T T, Ouyang X, Zheng J, et al. Structural analysis of the novel influenza A (H7N9) viral Neuraminidase interactions with current approved neuraminidase inhibitors Oseltamivir, Zanamivir, and Peramivir in the presence of mutation R289K[J]. BMC bioinformatics, 2013, 14(16): S7.

[8] Miotto O, Heiny A T, Tan T W, et al. Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis[J]. BMC bioinformatics, 2008, 9(1): S18.

[9] Chen H, Zhou X, Zheng J, et al. Rules of co-occurring mutations characterize the antigenic evolution of human influenza A/H3N2, A/H1N1 and B viruses[J]. BMC Medical Genomics, 2016, 9(3): 229.

[10] Nobusawa E, Aoyama T, Kato H, et al. Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses[J]. Virology, 1991, 182(2): 475-485.

[11] Burke D F, Smith D J. A recommended numbering scheme for influenza A HA subtypes[J]. PloS one, 2014, 9(11): e112302.

[12] Petersen T N, Brunak S, von Heijne G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions[J]. Nature methods, 2011, 8(10): 785-786.

[13] Hai-ting Z. Machine learning and bioinformatics[J]. Information and Control, 2003, 32(4): 352-7.

[14] Urbanowicz R J, Moore J H. Learning classifier systems: a complete introduction, review, and roadmap[J]. Journal of Artificial Evolution and Applications, 2009, 2009: 1.

[15] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update[J]. ACM SIGKDD explorations newsletter, 2009, 11(1): 10-18.

[16] Deem M W, Pan K. The epitope regions of H1-subtype influenza A, with application to vaccine efficacy[J]. Protein Engineering Design and Selection, 2009, 22(9): 543-546.

[17] Baigent S J, McCauley J W. Influenza type A in humans, mammals and birds: Determinants of virus virulence, host - range and interspecies transmission[J]. Bioessays, 2003, 25(7): 657-671.

[18] Kash J C, Qi L, Dugan V G, et al. Prior infection with classical swine H1N1 influenza viruses is associated with protective immunity to the 2009 pandemic H1N1 virus[J]. Influenza and other respiratory viruses, 2010, 4(3): 121-127.

- [19] Bragstad K, Nielsen L P, Fomsgaard A. The evolution of human influenza A viruses from 1999 to 2006: a complete genome study[J]. Virology journal, 2008, 5(1): 40.
- [20] Cho K J, Lee J H, Hong K W, et al. Insight into structural diversity of influenza virus haemagglutinin[J]. Journal of General Virology, 2013, 94(8): 1712-1722.
- [21] Piralla A, Pariani E, Rovida F, et al. Segregation of virulent influenza A (H1N1) variants in the lower respiratory tract of critically ill patients during the 2010–2011 seasonal epidemic[J]. PLoS One, 2011, 6(12): e28332.
- [22] Tewawong N, Prachayangprecha S, Vichiwattana P, et al. Assessing antigenic drift of seasonal influenza A (H3N2) and A (H1N1) pdm09 viruses[J]. PloS one, 2015, 10(10): e0139958.
- [23] Ginting T E, Shinya K, Kyan Y, et al. Amino acid changes in hemagglutinin contribute to the replication of oseltamivir-resistant H1N1 influenza viruses[J]. Journal of virology, 2012, 86(1): 121-127.

Columns on Last Page Should Be Made As Close As Possible to Equal Length

Authors' background

Your Name	Title*	Research Field	Personal website
YIN RUI	Phd candidate	Biomedical Informatics Machine learning	N/A
ZHOU XINRUI	Phd candidate	Biomedical Informatics Machine learning	N/A
Fransiskus Xaverius Ivan	Post Doc	Biomedical Informatics Computational Structural Biology	N/A
ZHENG JIE	Assistant Professor	Computational systems biology, Bioinformatics, computational genomics and epigenomics, Algorithm design, Scientific data mining and knowledge discovery	http://www.ntu.edu.sg/home/zhengjie/
VINCENT T.K. CHOW	Assoc Professor	Mouse models for understanding the molecular pathogenesis of influenza and pneumococcal pneumonia. Activation and formation of neutrophil extracellular traps in severe influenza and secondary pneumococcal pneumonia. Genomics and proteomics of emerging and re-emerging infectious diseases. Vaccine strategies, molecular surveillance and seroepidemiology of microbial infections	https://scholar.google.com/citations?user=7e7AUXkAAAAJ&hl=en
KWOH CHEE KEONG	Assoc Professor	Computational Structural Biology Ontology for Knowledge Representation Machine Learning and Statistical Inference	http://www.ntu.edu.sg/home/asckwoh/

*This form helps us to understand your paper better, **the form itself will not be published.**

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor