viruses

MDPI

Article

Evolutionary Insights from Association Rule Mining of Co-Occurring Mutations in Influenza Hemagglutinin and Neuraminidase

Valentina Galeone [1,2,†], Carol Lee [2], Michael T. Monaghan [3,4], Denis C. Bauer [2] and Laurence O. W. Wilson [2,5,*]

[1] Institute of Computer Science, Freie Universität Berlin, 14195 Berlin, Germany; valentina_galeone@outlook.it
[2] Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Sydney, NSW 2145, Australia; carol.lee@csiro.au (C.L.); denis.bauer@csiro.au (D.C.B.)
[3] Institute of Biology, Freie Universität Berlin, 14195 Berlin, Germany; michael.monaghan@igb-berlin.de
[4] Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), 12587 Berlin, Germany
[5] Department of Biomedical Sciences, Macquarie University, Sydney, NSW 2109, Australia
[*] Correspondence: laurence.wilson@csiro.au
[†] Current address: Genome Competence Center, Robert Koch Institute, 13353 Berlin, Germany.

Abstract: Seasonal influenza viruses continuously evolve via antigenic drift. This leads to recurring epidemics, globally significant mortality rates, and the need for annually updated vaccines. Co-occurring mutations in hemagglutinin (HA) and neuraminidase (NA) are suggested to have synergistic interactions where mutations can increase the chances of immune escape and viral fitness. Association rule mining was used to identify temporal relationships of co-occurring HA–NA mutations of influenza virus A/H3N2 and its role in antigenic evolution. A total of 64 clusters were found. These included well-known mutations responsible for antigenic drift, as well as previously undiscovered groups. A majority (41/64) were associated with known antigenic sites, and 38/64 involved mutations across both HA and NA. The emergence and disappearance of N-glycosylation sites in the pattern of N-X-[S/T] were also identified, which are crucial post-translational processes to maintain protein stability and functional balance (e.g., emergence of NA:339ASP and disappearance of HA:187ASP). Our study offers an alternative approach to the existing mutual-information and phylogenetic methods used to identify co-occurring mutations, enabling faster processing of large amounts of data. Our approach can facilitate the prediction of critical mutations given their occurrence in a previous season, facilitating vaccine development for the next flu season and leading to better preparation for future pandemics.

Keywords: influenza; H3N2; association rule mining; antigenic drift; co-occurring mutations

check for updates

Citation: Galeone, V.; Lee, C.; Monaghan, M.T.; Bauer, D.C.; Wilson, L.O.W. Evolutionary Insights from Association Rule Mining of Co-Occurring Mutations in Influenza Hemagglutinin and Neuraminidase. *Viruses* **2024**, *16*, 1515. https://doi.org/10.3390/v16101515

Academic Editor: Jianjun Chen

Received: 31 July 2024
Revised: 29 August 2024
Accepted: 11 September 2024
Published: 25 September 2024

1. Introduction

Seasonal influenza viruses (Orthomyxoviridae) are responsible for recurring epidemics worldwide, leading to approximately 250,000 to 500,000 deaths each year [1]. Most of these cases are caused by Influenza types A and B. The former circulates in animal hosts (bird and swine) and has caused devastating pandemics, for example, the Spanish Flu (1918) and swine flu (2009) caused by H1N1, and Hong Kong flu (1968) caused by H3N2 [2]. Influenza B primarily infects humans and has been circulated in human populations since the 1940s, having since diverged into two main lineages: B/Victoria and B/Yamagata [3]. Given their significant impact on human health, these two types are the primary targets for existing influenza vaccines. However, despite extensive research, the success of influenza can be attributed to its ongoing evolution and efficient transmission between hosts, allowing it to evade host immunity that results from previous infections or vaccinations [4].

The surface proteins hemagglutinin (HA) and neuraminidase (NA) play critical roles in viral replication and successful infection [5]. The receptor binding site (RBS) in the globular head domain of HA binds onto sialic acids (SA) on the surface of host cells, while NA is responsible for cleaving the HA–SA bond of budding virion for release and infection of new cells [6]. The gradual accumulation of mutations (antigenic drift) in the RBS, namely the five known antigenic regions, can influence host specificity and cell types [7–9], constantly challenging the effectiveness of new vaccines. However, the close proximity of NA and HA indicates that immune pressure caused by immunisation can generate favourable mutations in HA or NA to increase specificity and antigenicity or allow efficient release of virions, respectively [5,10,11]. This suggests that the co-evolution of HA and NA mutations leads to enhanced virus transmission and overall viral fitness and outbreaks. As seasonal influenza evolution involves simultaneous mutations, not just gradual single-point changes [12], a method to rapidly detect and analyse mutation groups, establish temporal relationships, and potentially uncover cause-and-effect links would be invaluable to identify important co-occurring mutations in influenza and discover potential functional links.

Many methods have been used to monitor mutation sites under positive selection-driving for or maintaining beneficial mutations, including statistical analysis and machine learning as an alternative to phylogenetics. Reconstructing evolutionary events through phylogenetics (maximum likelihood or Bayesian methods) often requires significant computational resources but allows for a more precise understanding of the chronological order of individual mutations. Association rule mining (ARM) offers a suitable alternative to phylogenetics or methods such as mutual information (MI), which only examines pairwise interactions [13–15]. This technique operates on transactional data: a "transaction" represents a set of items (mutations) that occur together frequently or are connected non-randomly and association rules that describe that the relationship between items are generated through the frequency of these itemsets [16,17]. ARM has been used to determine the various contribution of mutations to host range, pandemic/seasonal influenza, and the antigenic evolution of influenza virus [16,18,19] and has been applied to other pathogens and diseases as well [20–25]. The popularity and versatility of ARM arise from its capacity to uncover groups of key associations within datasets without demanding significant computational power, thanks to advancements in algorithmic efficiency [26,27]. Furthermore, the results are easily interpretable, making ARM an accessible and powerful tool for data analysis. The potential of this method, particularly when applied to co-occurring mutations in influenza, was first explored by Chen et al. (2016) [16].

Building on the work of [16] this study differs in three key approaches: (1) we limited the dataset to sequences from the year 2005 onwards. This decision was influenced by the increased availability of sequenced viruses in databases due to next generation sequencing (NGS) technologies. Chen et al. [16] collected data for the H3N2 subtype following the year 1968 (Hong Kong flu outbreak), which means that some years had a particularly small number of sequences. Therefore, limiting the range of years allows a notably higher number of sequences to be retained for study, enabling us to detect clusters of simultaneous mutations evolving exactly from one flu season to the next. (2) In our work, criteria used to identify sequences that are evolutionarily close were useful in excluding sequences that may not be genetically related. This enabled us to confidently assess which mutations occurred from one flu season to the next. (3) We included both HA and NA to detect co-evolving mutations, as these two proteins are closely interconnected in terms of function and evolution [28,29].

This study describes the application of ARM to detect co-occurring mutation clusters in the HA and NA of influenza virus A H3N2, combined with network analysis and phylogenetic analysis as a validation step by tracking the associations found by ARM. The aim of this study was to develop a method that can effectively identify patterns of co-evolution within the important HA and NA proteins of influenza A and establish links between functional mutations.

2. Materials and Methods

The datasets: Data were collected from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) [30]. The data used in this study included the complete sequences for

HA and NA of influenza virus A found within human hosts in North America. The scope was limited to North America to avoid potential biases in the representation of influenza virus strains, as most available sequences were sourced from this region. This encompassed 6915 sequences for the H3N2 subtype, spanning from 2006 to 2020. Sequences with >5% ambiguous characters (i.e., nucleotides other than A, C, G, and T) were removed and then organised into flu seasons based on whether they were collected before or after September 1st of a given year.

Data pre-processing: Additional filtering was performed to exclude potential reassorted sequences and retain sequences from H3N2 only. Sequences from consecutive bins were aligned using Clustal Omega, and the distance matrix was calculated using Kimura Two-Parameter (K2P) in the EMBOSS distmat command [31]. Preliminary results indicated that sequences from consecutive flu seasons typically ranged from 1–5 substitutions per 100 amino acids. BLAST [32] was used to confirm the subtype and year, and outliers (>5 substitutions) either matched a different subtype, host, or year contrary to metadata and were eliminated (0.1%) assuming misclassification or as an outcome of reassortment events.

The DNA sequences were translated into amino acid sequences, and where necessary, leading gaps were retrieved from NCBI and appended. Sequences with insertions were removed from the dataset (20 sequences; <0.1%), as our methods do not handle insertions. Custom Python scripts were used to process indels and duplications. The names of the selected strains are provided in Table S1. A summary of the sequences retained after each step is shown in Table 1.

Table 1. Total number of sequences after each step of the pre-processing workflow.

Step	H3N2
Download data from BV-BRC	13,543
Selecting the data bins	12,865
Removing incomplete sequences	12,622
Removing evolutionary distant sequences	12,612
Removing sequences with insertions	12,604
Removing duplicates	6915

Association rule mining: ARM was used to extract co-occurring mutations between and within the HA and NA glycoproteins and their transitions between consecutive flu seasons. A modified version of the association rule function from mlxtend version 0.21.0 [33] was used to identify itemsets (specific combinations of mutations between flu seasons). This was to generate rules associated with antigenic sites for HA and NA [7,34]. To compute the complete set of all frequent itemsets, the algorithm FPgrowth was used. Default values were used for the calculation of frequent itemsets (minimum support value = 0.05) and association rules (minimum confidence = 0.5). In practical terms, this meant that all rules were generated with a confidence level higher than 50%. Default values were relatively lenient to allow the generation of sufficient rules and order by the highest confidence.

This process was repeated multiple times to collect N number of transactions. A value of N = 250 performed well, yielding consistent results across different random seeds while maintaining a low probability of randomly selecting the same pair of flu sequences and introducing bias. This formed the basis of downstream analysis to identify patterns of co-occurring mutations over consecutive flu seasons. This final approach involved extracting transactions independently from each pair of flu seasons, characterising these transitions and detecting recurring patterns by comparing the clusters across years. The number of sequences available prior to 2006 at the time of collection consisted of <500 sequences per season. Thus, flu sequences before 2006 were not considered to ensure that the categories included were sufficiently large to reduce the risk of resampling bias when increasing the number of selected sequences. Several metrics were used to measure the strength and confidence of each association (Table S2). We included an additional metric, Zhang's metric, for a more precise measure of confidence and an extension of Lift [35]. The scale ranged from −1 (complete disassociation) to 1 (complete association). Association rules with a Zhang's metric > 0.85 were prioritised to focus on robust evidence of association.

To address false or reversed mutations, we assumed that clusters of mutations occur in a single direction (from one flu season to the next, not vice versa) based on the year. Two methods were employed to clarify the order of mutations: (1) phylogenetic analysis, in which we examined the order of mutations to determine their sequential occurrence, and (2) frequency plots, where mutations with the highest frequencies were considered valid, while those appearing in the opposite direction were considered invalid (see Figure S1).

Relevant associations detected between the two flu seasons were visualised through a network displaying co-occurring mutations using Networkx (version 3.1) [36] and Pyvis (version 0.3.1) [37].

Sequence variability and phylogenetic validation of mutation transitions: The maximum likelihood (ML) in IQ-TREE [38] was used to construct separate phylogenetic trees for H3N2 hemagglutinin and neuraminidase and establish a mutation threshold for our approach. The FLU amino acid substitution model [39] with the Gamma model was used to account for rate heterogeneity. This model specifically addresses the evolution of influenza virus sequences. The threshold was to exclude pairs belonging to distinct clades of the same influenza type that are not likely to have occurred simultaneously in one year and, thus, discard those exceeding the threshold. In certain years, such as 2012 and 2019, the division in subclades is more evident. A threshold of mutations T = 15 was chosen that was suitable to avoid the alignment of two distantly related sequences that belong to different subclades (Figure S2). Additionally, tree data from IQ-TREE was used to map mutation clusters to validate and cross-reference the identified clusters.

Shannon's entropy and frequency plots: Shannon's entropy serves as a valuable tool for assessing the variability of amino acid positions within multiple sequence alignment. Multiple sequence alignments for each flu season were used to calculate the entropy values for each position, and the mean value across bins was used to assess the overall variability. Positions with a high average entropy value were indicative of the positions having undergone multiple changes over time, providing valuable insights into protein evolution. Entropy was calculated using the following formula:

$$H = -\sum_{i=1}^{N} p(x_i)\log_2(p(x_i)).$$

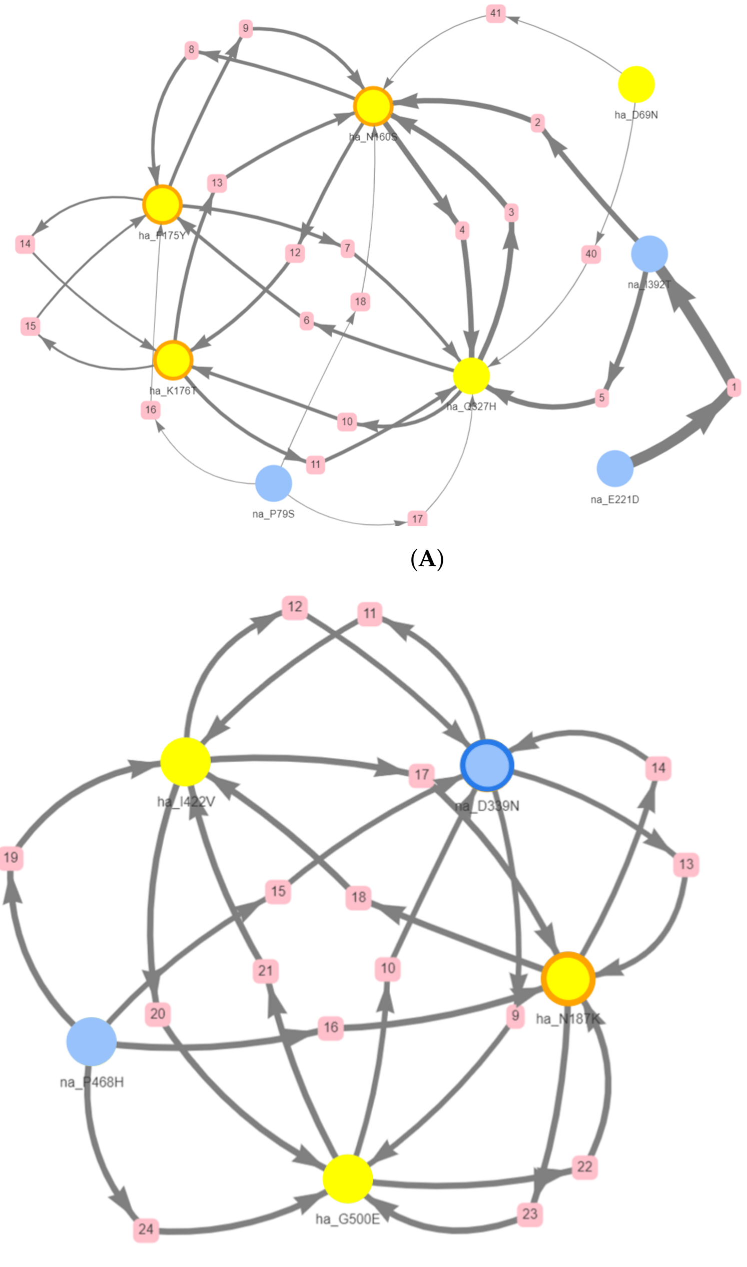
Frequency plots were used to visualise the amino acid frequencies of positions with the highest entropy over time. Transitions in the same position with the lowest frequency were excluded from the results, and no association rules were generated.

3. Results

3.1. Association Rules and Clusters of Mutations

A total of 1647 rules of association in H3N2, spanning from 2006 to 2020, with confidence larger than 0.5 and support larger than 0.05, were identified. Further filtering by applying a stringent threshold of Zhang's metrics > 0.85 was used to identify rules with stronger evidence of association. A total of 64 clusters of mutations were found, representing small sets of mutations ranging from two to seven residues, which evolved together from one flu season to the next, with a mean of about four to five clusters occurring for each transition (Table S3). The majority of clusters found in the H3N2 dataset included amino acids at antigenic sites (41 clusters out of 64), possibly linking amino acids involved in antigenic variation with more distant positions. Furthermore, the clusters comprise mutations found in hemagglutinin and neuraminidase or spanning across both proteins.

We visualised the clusters of mutations that included these positions as networks (Figure 1). Each pink box is a (directed) rule identified by a number. These mutation networks not only revealed associations among mutations but also drew attention to noteworthy instances. Notably, we readily identified connections between mutations in the two proteins and specifically within antigenic sites. Figure 1A illustrates the mutation site ha_160 playing a central role in the network and formed connections with HA mutations at positions 69, 175, 176, and 327 and NA mutations at positions 79 and 1392, except for 221, which is individually associated with na_1392. The four HA mutations (160, 175, 176, and 327) are interconnected bidirectionally, suggesting possibilities of A to B and B to A transitions. However, the unidirectional na_I392T to ha_N160S suggests the occurrence of N160S when 392T occurs but not the reverse. Figure 1B shows a similar cluster where antigenic site 339 is interconnected with four other positions.



(**A**)

(**B**)

Figure 1. Cluster of mutations in H3N2 during the transition from the (**A**) 2012/13 to 2013/14 and (**B**) 2014/15 to 2015/16 flu season. Transitions include HA and NA antigenic sites. Filled circles indicate mutations in hemagglutinin (yellow) and neuraminidase (blue), and an orange or dark blue border (e.g., node ha_D339N) indicates that the mutation occurred at an antigenic site, pink boxes indicates a (directed) rule identified by a number. The thickness of the edges represents the support of the rule and the direction show the antecedent and consequent.

3.2. Shannon's Entropy and Frequency Plots

Positions with a high average entropy value are indicative of having undergone multiple changes over time, providing valuable insights into protein evolution. For HA, the three positions with the highest entropy were identified as 144, 158, and 160 (Figure S3). These positions are all located within the antigenic regions, and position 158, the one with the highest entropy in the H3N2 dataset, is among the seven key amino acid sites responsible for driving antigenic changes [40]. Phylogenetic analysis (Figure 2) and frequency plots (Figure S4) indicated positions 158 and 144 were found to be strongly associated during both the transition from the flu season 2011/12 to 2012/13 and from 2012/13 to 2013/14 with our method (Figure S3). In both cases, they were not associated with other mutations but formed a cluster consisting of only two elements. However, when examining positions 160 and 144, no clusters demonstrated their association, which was verified with the HA phylogenetic tree (Figure 2). Phylogenetic analysis also showed incongruence between HA and NA, which can be attributed to reassortment events (Figure S5). Notably, the presence of subclades was clear, with hemagglutinin displaying more pronounced subclade formations. In specific cases, such as the years 2012 and 2019, the division into distinct clades was more evident.

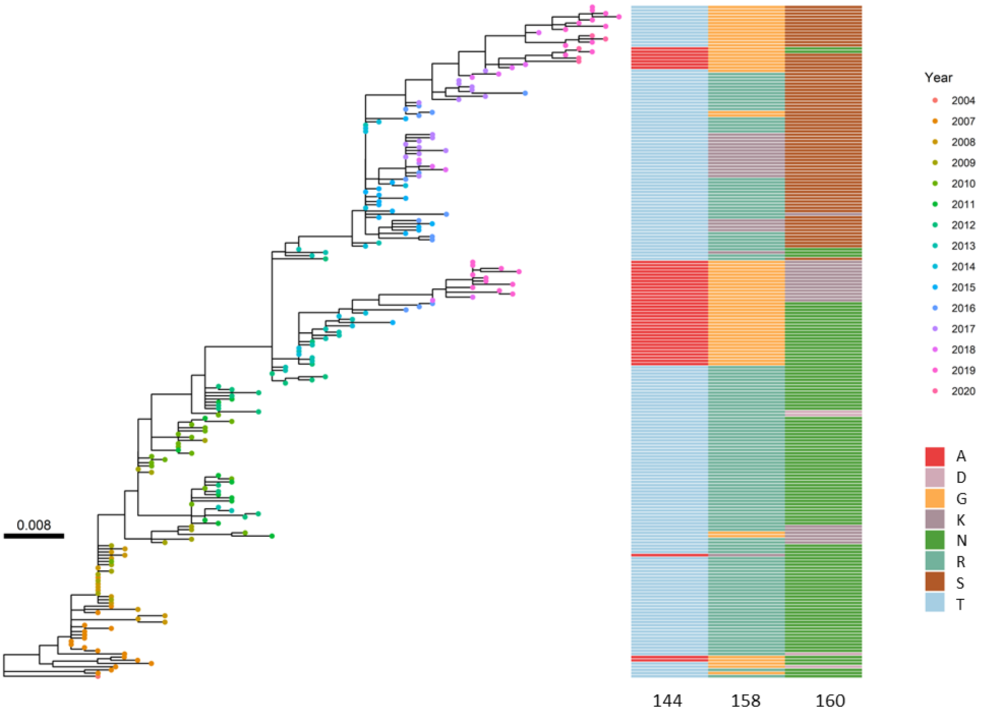


Figure 2. Maximum likelihood phylogenetic tree of 15 randomly selected HA sequences from each flu season with the amino acid illustrated for positions 144, 158, and 160 for each sequence. Coloured dots indicate the year isolated, showing highly temporal relationships between sequences and the protein position with the highest entropy.

3.3. N-Glycosylation Sites

An important feature that appears from the results is the emerging and disappearing of N-glycosylation sites, which are crucial post-translational processes that impact the protein's stability and function by attaching sugar molecules, thereby influencing its biological activity and interactions. For example, in Figure 1B, the two positions in the antigenic sites provided evidence of the emergence of asparagine at position 339 of NA, associated with the disappearance of asparagine at position 187 of HA. The presence of asparagine at these positions suggests the potential formation of N-glycosylation sites. It is noteworthy that among the clusters, 32 out of 64 contain at least one instance of either an emerging or disappearing asparagine. The case of na_D339N and ha_N189K is not unique; in many other instances, an emerging asparagine is coupled with a disappearing one. Another interesting pattern is the emergence of the sequence pattern N-X-[S/T], as in the highlighted cluster occurring during the transition 13/14–14/15, which includes na_S247T and na_S245N. These two mutations have been observed to co-evolve, leading to the creation of an N-glycosylation site at position 245. We can hypothesize that a similar mechanism took place in the clusters listed in Table 2 (full list of clusters in Table S3) that contain mutations na_N465S and na_D463N during the transition 18/19–19/20. In this case, it is reasonable to conclude that mutations 465 and 463 co-evolved, resulting in the formation of a new N-glycosylation site at position 463. Furthermore, the emergence of this new N-glycosylation site is coupled with the potential loss of a site at position 110 of HA.

Table 2. Subset of co-occurring mutation clusters in H3N2. The asterisks (*) indicate that the mutation occurred at an antigenic site.

Flu Season	HA Mutations	NA Mutations
07/08–08/09	N160*K , N205K, V229*A, K174*N, E78*K	-
12/13–13/14	F175*Y, Q327H, D69N, K176*T, N160*S	E221D, P79S, I392T
13/14–14/15	-	S247T, S245N
15/16–16/17	I422V, G500E, N187*K	N339*N
18/19–19/20	K99E, I538M, Y110N	N465S, D463N, G346*D

4. Discussion

Research has demonstrated that the antigenic drift of seasonal influenza viruses is not solely driven by gradual single-point mutations but also by simultaneous mutations [12,41]. For this reason, there is a need for methods that are capable of rapidly detecting and analyzing co-occurring groups of mutations, identifying temporal relationships within such groups, reconstructing the order of events underlying major evolutionary changes, and eventually uncovering any cause–effect relationships that may exist among these mutations. These data can be used to establish effective predictive methods for monitoring the emergence of new viral strains that could be more virulent or influence current vaccination protocols. The current study presents a dedicated approach designed to address this initial step by rapidly characterizing groups of simultaneous mutations through the application of association rule mining principles.

Several clusters of co-occurring mutations were found to extend across both hemagglutinin and neuraminidase, suggesting interconnected functionalities between these proteins, a hypothesis that should be better investigated to identify its potential roles in influenza pathogenicity. HA and NA work in tandem to ensure efficient virion release for further infection of host cells [42]. Many of the association rules involved both HA and NA (38/64) and include one to five HA mutations and one to three NA mutations. The functional balance between HA and NA proteins needs to be maintained due to their complementary functions, where the evolutionary potential of HA is influenced by NA in an effort to increase viral fitness through immune escape [5,43]. Mutations in NA can be restricted so as not to impact the epitope binding potential of HA for initial infection, not dissimilar to the non-random reassortment of specific HA and NA subtypes for cross-species infection [5,18]. These co-occurring mutations in both HA and NA further provide insights into NA-independent resistance, where many NA inhibitor-resistant mutants are present due to mutations in HA through reduced binding affinity and reducing the dependency on NA for virion release [44]. This dependency may indicate why NA inhibitor resistance is less prevalent than adamantane (1–4% adults shedding resistant virus versus up to 23%, respectively) [45,46], which targets the M2 protein [47]. The current study did not identify any NA inhibitor-resistant mutations in either HA or NA. However, further studies may benefit from categorising influenza sequences by their antibody affinity to identify rules associated with antiviral resistance.

We also identified clusters linked to the emergence or disappearance of N-glycosylation sites, shedding light on glycosylation-related changes in protein function. The glycosylation of HA and NA is indicative of immune evasion without loss of viral fitness [48]. Our results also highlighted the na_S247T and na_S245N mutation observed circa 2015 with reduced NA antibody binding [49]. This is in concordance with the N-X-[S/T] pattern observed to prevent antibody contact with underlying residues and, thus, impacts vaccine efficacy [50]. The HA and NA proteins function synergistically to successfully infect host cells, and modifications to HA–NA, such as through glycosylation, can impact viral fitness. Our study identified several mutational transitions in the years from 2015–2020, which we found in IAV subtypes that differed from the vaccine strains (A/Singapore/infimh-16-0019/2016). These co-occurring mutations include na_P126L, na_K220N, and na_V303I [51–53], in addition to clustering with ha_E78G, ha_K108R, ha_T151K, ha_R158G, and ha_H327Q, which differ from the A/Hong Kong/4801/2014(H3N2) vaccine strain. Interestingly, we did not find na_X329N and na_E344K to be co-occurring, which is often linked to higher neuraminidase-inhibiting (NI) activity. However, this may be due to the decreased percentage of isolates with N-glycosylation at na_329 since 2015 [54]. We also noted the na_L140I and na_V149A mutations differing from the A/Switzerland/8060/2017 vaccine strain used for the southern hemisphere. The latter mutation was also close to the active site and may have influenced sialidase activity [55]. Additionally, we found notable clusters resulting in the loss of glycosylation sites such as ha_N187K from 2014. This co-occurred with other HA mutations (ha_I422V and ha_G500E) commonly found around 2016–2017 [52] and na_P468H and na_339N, with the latter being an emerging glycosylation site. HA and NA have a complex co-evolution dynamic, constantly changing to modulate binding and cleaving activities and have the potential to compensate for function in the other protein [56]. ARM has the potential to extract these complex relationships and identify these frequently interacting sites. The potential of identifying mutations contributing to glycosylation or sequons and evaluating their influence on antibody binding and vaccine efficacy would improve influenza vaccine development through the optimisation of using both HA and NA mutations for consideration.

In the current study, the limited 15-year range excluded insights into mutations that may have occurred multiple times over the earliest sequences available. As a result, only a few mutations recurred in different combinations within the database, with no clear pair or group of mutations exhibiting repeated occurrences. The inclusion of earlier sequences (<2006) with consideration of potential resampling bias could offer insights into more historical trends. Nonetheless, a noticeable pattern that emerged was the consistent appearance and disappearance of asparagine, which potentially represents the emergence and disappearance of N-glycosylation sites. These mutations often occurred in pairs: an N appearing in a new position was coupled with an N disappearing from another position, spanning both the HA and NA proteins.

Association rule mining is a powerful tool to rapidly detect association in a transaction dataset: the efficiency is given by a runtime of less than one minute. Correctly characterising linked mutations and identifying the major determinants that drive their associations is the first critical step in developing an effective tool that can prepare for future pandemics by detecting key groups of associated mutations in time. In addition to ARM, a comprehensive predictive tool for monitoring virus evolution and anticipating future mutations should integrate information from various sources, not solely relying on association rule mining. *In silico* modelling (e.g., AlphaFold [57]) of mutational transitions identified here can provide further insight into the impact of these mutations on the structure of the proteins and the potential effect on receptor binding and cleavage. Another avenue of exploration is the incorporation of phylogenetic analysis as an integral component rather than using it solely for validation, as was the case here. Such an approach would allow further valuable insights into the evolutionary history of key mutations, particularly those within antigenic sites or receptor-binding sites. By tracking the lineage of these mutations on a phylogenetic tree, deeper insights into their emergence and persistence over time will complement the information obtained through association rule mining regarding more distant mutations associated with these key positions.

5. Conclusions

These findings highlight the potential of ARM to identify co-occurring mutations of functional interest. ARM provides a valuable foundation for further analysis and the potential development of predictive tools. ARM can be extended to other influenza subtypes to uncover broader evolutionary patterns and co-occurring mutations that may be implicated in preparedness for future outbreaks and be further developed with predictive algorithms.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/v16101515/s1, Table S1: Name of strains used in this study; Table S2: Description and formula of metrics used to evaluate the association rules generated in this study; Table S3: Summary of the 64 co-occurring mutation clusters in H3N2. Figure S1: Illustration of potential 'reverse mutations', where sequence 1 and 2 are randomly selected from two consecutive flu seasons (bins) with sequence 2 from a more recent bin; Figure S2: Boxplot showing the different number of mutations when the sequences belong to the same clade compared to belonging to different clades in (a) 2012 and (b) 2019; Figure S3: Positions with a high entropy showed a larger number of amino acids in the frequency plots; Figure S4: Frequency plots of amino acid positions depicting residue frequency for each flu season; Figure S5: H3N2 maximum likelihood phylogenetic trees generated with IQ-tree, employing 15 randomly selected sequences from each flu season bin.

Author Contributions: Conceptualization, C.L. and L.O.W.W.; Formal analysis, V.G.; Investigation, V.G.; Methodology, V.G.; Supervision, C.L.; Visualization, V.G.; Writing—original draft, V.G. and C.L.; Writing— review and editing, V.G., C.L., M.T.M., D.C.B. and L.O.W.W. All authors have read and agreed to the published version of the manuscript.

Funding: V.G. would like to acknowledge the DAAD scholarship "Internationale Forschungsaufenthalte für Informatikerinnen und Informatiker (IFI)" for supporting her work in CSIRO, Australia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All custom Python scripts for the ARM analyses are available in Github: https://github.com/valegale/influenza_mutations (v1.0.0), and data pre-processing scripts can be found at https://github.com/valegale/preprocessing_influenza (v1.0.0).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARM	Association Rule Mining
BLAST	Basic Local Alignment Search Tool
BV-BRC	Bacterial and Viral Bioinformatics Resource Center
DNA	Deoxyribonucleic Acid
EMBOSS	The European Molecular Biology Open Software Suite
HA	Hemaglutinin
IAV	Influenza A Virus
MI	Mutual Information
ML	Maximum Likelihood
NA	Neuraminidse
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
RBS	Receptor Binding Site
SA	Sialic Acids

References

1. Iuliano, A.D.; Roguski, K.M.; Chang, H.H.; Muscatello, D.J.; Palekar, R.; Tempia, S.; Cohen, C.; Gran, J.M.; Schanzer, D.; Cowling, B.J.; et al. Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. *Lancet* **2018**, *391*, 1285–1300 [CrossRef] [PubMed]
2. Kumar, B.; Asha, K.; Khanna, M.; Ronsard, L.; Meseko, C.A.; Sanicas, M. The emerging influenza virus threat: Status and new prospects for its therapy and control. *Arch. Virol.* **2018**, *163*, 831–844. [CrossRef] [PubMed]
3. Shao, W.; Li, X.; Goraya, M.U.; Wang, S.; Chen, J.-L. Evolution of Influenza A Virus by Mutation and Re-Assortment. *Int. J. Mol. Sci.* **2017**, *18*, 1650. [CrossRef] [PubMed]
4. Petrova, V.N.; Russell, C.A. The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* **2018**, *16*, 47–60. [CrossRef]
5. Kosik, I.; Yewdell, J.W. Influenza Hemagglutinin and Neuraminidase: Yin–Yang Proteins Coevolving to Thwart Immunity. *Viruses* **2019**, *11*, 346. [CrossRef]
6. Shtyrya, Y.A.; Mochalova, L.V.; Bovin, N.V. Influenza virus neuraminidase: Structure and function. *Acta Naturae* **2009**, *1*, 26–32. [CrossRef]
7. Lee, M.-S.; Chen, J.S.-E. Predicting Antigenic Variants of Influenza A/H3N2 Viruses. *Emerg. Infect. Dis.* **2004**, *10*, 1385–1390. [CrossRef]
8. Mair, C.M.; Ludwig, K.; Herrmann, A.; Sieben, C. Receptor binding and pH stability—How influenza A virus hemagglutinin affects host-specific virus infection. *Biochim. Biophys. Acta.* **2014**, *1838*, 1153–1168 [CrossRef]
9. Yang, Z.-Y.; Wei, C.-J.; Kong, W.-P.; Wu, L.; Xu, L.; Smith, D. F.; Nabel, G. J. Immunization by avian H5 influenza hemagglutinin mutants with altered receptor binding specificity. *Science* **2007**, *317*, 825–828. [CrossRef]
10. lyushina, N.A.; Komatsu, T.E.; Ince, W.L.; Donaldson, E.F.; Lee, N.; O’Rear, J.J.; Donnelly, R.P. Influenza A virus hemagglutinin mutations associated with use of neuraminidase inhibitors correlate with decreased inhibition by anti-influenza antibodies.*Virol. J.* **2019**, *16*, 149. [CrossRef]
11. Wang, F.; Wan, Z.; Wang, Y.; Wu, J.; Fu, H.; Gao, W.; Shao, H.; Qian, K.; Ye, J.; Qin, A. Identification of Hemagglutinin Mutations Caused by Neuraminidase Antibody Pressure. *Microbiol. Spectr.* **2021**, *9*, e01439-21. [CrossRef] [PubMed]
12. Shih, A.C.-C.; Hsiao, T.-C.; Ho, M.-S.; Li, W.-H. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 6283–6288. [CrossRef] [PubMed]
13. Arcos, S.; Han, A.X.; Te Velthuis, A.J.W.; Russell, C.A.; Lauring, A.S. Mutual information networks reveal evolutionary relationships within the influenza A virus polymerase. *Virus Evol.* **2023**, *9*, vead037. [CrossRef] [PubMed]
14. Gong, Y.-N.; Chen, G.-W.; Suchard, M.A. A novel empirical mutual information approach to identify co-evolving amino acid positions of influenza A viruses. *Comput. Biol. Chem.* **2012**, *39*, 20–28. [CrossRef]
15. Xia, Z.; Jin, G.; Zhu, J.; Zhou, R. Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. *Bioinformatics* **2009**, *25*, 2309–2317. [CrossRef]
16. Chen, H.; Zhou, X.; Zheng, J.; Kwoh, C.-K. Rules of co-occurring mutations characterize the antigenic evolution of human influenza A/H3N2, A/H1N1 and B viruses. *BMC Med. Genom.* **2016**, *9*, 69. [CrossRef]
17. Kaur, M.; Kang, S. Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Comput. Sci.* **2016**, 85, 78–85. [CrossRef]
18. Kargarfard, F.; Sami, A.; Ebrahimie, E. Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm. *J. Biomed. Inform.* **2015**, *57*, 181–188. [CrossRef]
19. Kargarfard, F.; Sami, A.; Mohammadi-Dehcheshmeh, M.; Ebrahimie, E. 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 Genom.* **2016**, *17*, 925. [CrossRef]
20. Tandan, M.; Acharya, Y.; Pokharel, S.; Timilsina, M. Discovering symptom patterns of COVID-19 patients using association rule mining. *Comput. Biol. Med.* **2021**, *131*, 104249. [CrossRef]
21. Greenbaum, B.D.; Ghedin, E. Viral evolution: Beyond drift and shift. *Curr. Opin. Microbiol.* **2015**, *26*, 109–115. [CrossRef]
22. Indhumathy, M.; Nabhan, A.R.; Arumugam, S. A Weighted Association Rule Mining Method for Predicting HCV-Human Protein Interactions. *Curr. Bioinform.* **2018**, *13*, 73–84. [CrossRef]
23. Leung, K.S.; Lee, K.H.; Wang, J.F.; Ng, E.Y.; Chan, H.L.; Tsui, S.K.; Mok, T.S.; Tse, P.C.; Sung, J.J. Data mining on DNA sequences of hepatitis B virus. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8*, 428–440. [CrossRef] [PubMed]
24. Liang, B.; Li, X.; Zhang, Z.; Wu, C.; Liu, X.; Zheng, Y. Multidrug resistance analysis method for pathogens of cow mastitis based on weighted-association rule mining and similarity comparison. *Comput. Electron. Agric.* **2021**, *190*, 106411. [CrossRef]
25. Gakii, C.; Rimiru, R. Identification of cancer related genes using feature selection and association rule mining. *Inform. Med. Unlocked.* **2021**, *24*, 100595. [CrossRef]
26. Han, J.; Pei, J.; Yin, Y. Mining frequent patterns without candidate generation. *Data Min. Knowl. Discov.* **2000**, *29*, 1–12.
27. Zaki, M.J.; Gouda, K. Fast vertical mining using diffsets. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2003; pp. 326–335.
28. Jang, J.; Bae, S.-E. Comparative Co-Evolution Analysis Between the HA and NA Genes of Influenza A Virus. *Virology* **2018**, *9*, 1178122X1878832. [CrossRef]
29. Zeller, M.A.; Chang, J.; Vincent, A.L.; Gauger, P.C.; Anderson, T.K. Spatial and temporal coevolution of N2 neuraminidase and H1 and H3 hemagglutinin genes of influenza A virus in US swine. *Virus Evol.* **2021**, *7*, veab090. [CrossRef]

30. Olson, R.D.; Assaf, R.; Brettin, T.; Conrad, N.; Cucinell, C.; Davis, J.J.; Dempsey, D.M.; Dickerman, A.; Dietrich, E.M.; Kenyon, R.W.; et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): A resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* **2023**, *51*, D678–D689. [CrossRef]
31. Madeira, F.; Pearce, M.; Tivey, A.R.N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **2022**, *50*, W276–W279. [CrossRef]
32. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef] [PubMed]
33. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J. Open Source Softw.* **2018**, *3*, 638. [CrossRef]
34. Tulip, W.R.; Varghese, J.N.; Baker, A.T.; Van Donkelaar, A.; Laver, W.G.; Webster, R.G.; Colman, P.M. Refined atomic structures of N9 subtype influenza virus neuraminidase and escape mutants. *J. Mol. Biol.* **1991**, *221*, 487–497. [CrossRef] [PubMed]
35. Zhang, T. Association Rules. In Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications (PADKK '00), Kyoto, Japan, 18–20 April 2000; Springer: Berlin/Heidelberg, Germany, 2000; pp. 245–256.
36. Hagberg, A.A.; Schult D.A.; Swart, P.J. Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference SciPy2008, Pasadena, CA, USA, 19–24 August 2008; pp. 11–15.
37. Perrone, G.; Unpingco, J.; Lu, H. Network visualizations with Pyvis and VisJS. *arXiv* **2020**, arXiv:2006.04951. [CrossRef]
38. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
39. Dang, C.C.; Le, Q.S.; Gascuel, O.; Le, V.S. FLU, an amino acid substitution model for influenza proteins. *BMC Ecol. Evol.* **2010**, *10*, 99. [CrossRef]
40. Koel, B.F.; Burke, D.F.; Bestebroer, T.M.; Van Der Vliet, S.; Zondag, G.C.M.; Vervaet, G.; Skepner, E.; Lewis, N.S.; Spronken, M.I.J.; Russell, C.A.; et al. Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *IScience* **2013**, *342*, 976–979. [CrossRef]
41. Tria, F.; Pompei, S.; Loreto, V. Dynamically correlated mutations drive human Influenza A evolution. *Sci. Rep.* **2013**, *3*, 2705. [CrossRef]
42. Dou, D.; Revol, R.; Östbye, H.; Wang, H.; Daniels, R. Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement. *Front. immunol.* **2018**, *9*, 1581. [CrossRef]
43. Liu, T.; Wang, Y.; Tan, T.J.C.; Wu, N.C.; Brooke, C.B. The evolutionary potential of influenza A virus hemagglutinin is highly constrained by epistatic interactions with neuraminidase. *Cell Host Microbe* **2022**, *30*, 1363–1369.e4. [CrossRef]
44. Hurt, A.C.; Ho, H.-T.; Barr, I. Resistance to anti-influenza drugs: Adamantanes and neuraminidase inhibitors. *Expert Rev. Anti-infect. Ther.* **2006**, *4*, 795–805. [CrossRef] [PubMed]
45. Bright, R.A.; Medina, M.; Xu, X.; Perez-Oronoz, G.; Wallis, T.R.; Davis, X.M.; Povinelli, L.; Cox, N.J.; Klimov, A.I. Incidence of adamantane resistance among influenza A (H3N2) viruses isolated worldwide from 1994 to 2005: A cause for concern. *Lancet* **2005**, *366*, 1175–1181. [CrossRef] [PubMed]
46. Gubareva, L.V.; Kaiser, L.; Matrosovich, M.N.; Soo-Hoo, Y.; Hayden, F.G. Selection of Influenza Virus Mutants in Experimentally Infected Volunteers Treated with Oseltamivir. *J. Infect. Dis.* **2005**, *183*, 523–531. [CrossRef] [PubMed]
47. Nelson, M.I.; Simonsen, L.; Viboud, C.; Miller, M.A.; Holmes, E.C. The Origin and Global Emergence of Adamantane Resistant A/H3N2 Influenza Viruses. *Virology* **2009**, *388*, 270–78. [CrossRef] [PubMed]
48. Kim, H.; Webster, R.G.; Webby, R.J. Influenza Virus: Dealing with a Drifting and Shifting Pathogen. *Viral Immunol.* **2018**, *31*, 174–183. [CrossRef]
49. Wan, H.; Gao, J.; Yang, H.; Yang, S.; Harvey, R.; Chen, Y.-Q.; Zheng, N.-Y.; Chang, J.; Carney, P. J.; Li, X.; et al. The neuraminidase of A(H3N2) influenza viruses circulating since 2016 is antigenically distinct from the A/Hong Kong/4801/2014 vaccine strain. *Nat. Microbiol.* **2019**, *4*, 2216–2225. [CrossRef]
50. Chang, D.; Zaia, J. Why Glycosylation Matters in Building a Better Flu Vaccine. *Mol. Cell. Proteom.* **2019**, *18*, 2348–2358. [CrossRef]
51. Dudin, G.A.; Aziz, I.M.; Alzayed, R.M.; Ahmed, A.; Hussain, T.; Somily, A.M.; Alsaadi, M.M.; Almajhdi, F.N. Genetic Diversity and Evolutionary Kinetics of Influenza A Virus H3N2 Subtypes Circulating in Riyadh, Saudi Arabia. *Vaccines* **2023**, *11*, 702. [CrossRef]
52. Phyu, W.W.; Saito, R.; Kyaw, Y.; Lin, N.; Win, S.M.K.; Win, N.C.; Ja, L.D.; Htwe, K.T.Z.; Aung, T.Z.; Tin, H.H.; et al. Evolutionary Dynamics of Whole-Genome Influenza A/H3N2 Viruses Isolated in Myanmar from 2015 to 2019. *Viruses* **2023**, *14*, 2414. [CrossRef]
53. Boonnak, K.; Mansanguan, C.; Schuerch, D.; Boonyuen, U.; Lerdsamran, H.; Jiamsomboon, K.; Sae Wang, F.; Huntrup, A.; Prasertsopon, J.; Kosoltanapiwat, N.; et al. Molecular Characterization of Seasonal Influenza A and B from Hospitalized Patients in Thailand in 2018–2019. *Viruses* **2021**, *13*, 977. [CrossRef]
54. Ge, J.; Lin, X.; Guo, J.; Liu, L.; Li, Z.; Lan, Y.; Liu, L.; Guo, J.; Lu, J.; Huang, W.; et al. The Antibody Response Against Neuraminidase in Human Influenza A (H3N2) Virus Infections During 2018/2019 Flu Season: Focusing on the Epitopes of 329-N-Glycosylation and E344 in N2. *Front. Microbiol.* **2022**, *13*, 845088. [CrossRef] [PubMed]
55. Simon, B.; Pichon, M.; Valette, M.; Burfin, G.; Richard, M.; Lina, B.; Josset, L. Whole Genome Sequencing of A(H3N2) Influenza Viruses Reveals Variants Associated with Severity during the 2016–2017 Season. *Viruses* **2019**, *11*, 108. [CrossRef] [PubMed]

56. Mitnaul, L.J.; Matrosovich, M.N.; Castrucci, M.R.; Tuzikov, A.B.; Bovin, N.V.; Kobasa, D.; Kawaoka, Y. Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. *J. Virol.*, **2000**, *74*, 6015–6020. [CrossRef] [PubMed]
57. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.