NETWORK ANALYSIS OF CORRELATED MUTATIONS IN INFLUENZA

by

Uday Yallapragada
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

_________________________________ Dr. Iosif Vaisman, Dissertation Director

_________________________________ Dr. Dmitri Klimov, Committee Member

_________________________________ Dr. Ramin Hakami, Committee Member

_________________________________ Dr. Iosif Vaisman, Acting Director, School of Systems Biology

_________________________________ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

_________________________________ Dr. Peggy Agouris, Dean, College of Science

Date: ___________________________ Summer Semester 2017
George Mason University
Fairfax, VA

Network Analysis of Correlated Mutations in Influenza

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Uday Yallapragada
Master of Science
Michigan State University, 1998
Bachelor of Engineering
GITAM University, 1996

Director: Iosif Vaisman, Professor
Department of Bioinformatics and Computational Biology

Summer Semester 2017
George Mason University
Fairfax, VA

DEDICATION

I dedicate this dissertation to my parents for their love and support.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude for my advisor Dr. Vaisman for his guidance, patience and mentoring. He was very flexible and spent several hours striking insightful discussions. He has constantly challenged me and provided prudent advice. I was fortunate to be his student. I would like to thank my committee members Dr. Klimov and Dr. Hakami for their guidance. I would also like to thank Ms. Diane St. Germain for her administrative support and Mr. Chris Ryan for his help with infrastructure issues.

On a personal front, I would like to profoundly thank my wife Meera. I would not have completed this work if not for her patience, understanding and support. I also thank our children Vinay and Kavya for giving me the strength and determination to continue this journey. My sincere thanks also go out to my siblings and friends for their support and encouragement.

TABLE OF CONTENTS

List of Tables ix
List of Figures x
List of Abbreviations xiii
Abstract xiv
Chapter 1 - Introduction 1
  Dynamic Nature of Influenza A Virus 1
  Correlated Mutations 1
    Correlated Mutation Analysis in IAV Sequences 2
    Network Analysis 3
  Significance 3
    Antiviral Drugs 4
    Vaccines 4
    Epidemiology 5
  Objectives 5
    General properties of the networks of correlated mutations 6
    Subtype dependence of the network properties 6
    Protein structure and function implications of network features 7
    Computational contributions 7
  Dissertation Structure 8
Chapter 2 - Additional Background 10
  Influenza 10
  Classification of Influenza Virus 10
  Nomenclature of Influenza Strains 11
  Proteins in Influenza 12
  Structure of Influenza 13
  Life Cycle of Influenza 15
  Network Analysis 17
  Correlation Measures 19
    Maximal Information Coefficient 20

Chapter 3 – Classification of Influenza Sequences 21
Introduction 21
Background 22
Methods 26
Data 26
Pipeline 27
N-grams and feature extraction 28
Automated classification 28
Correlations 28
Classification Results 29
Drug Resistance 30
Viral Subtypes 36
Geography 37
Flu season 37
Similarity to 2009 pH1N1 38
Classification of DNA sequences 39
Classification of protein sequences using a reduced 3-letter alphabet 40
Correlations 40
Discussion 42
Chapter 4 - Literature Review 43
Correlated Mutations 43
Network Analysis 46
Chapter 5 - DataSets 49
Workflow 49
Details of Datasets 52
Chapter 6 - Computational Tools and Libraries 54
Chapter 7 - Implementation Details 58
Core Themes 58
Modularity 58
Best of breed 58
Visualizations 59
Extensibility 59

Graph Database ... 59
Sequence Encoding ... 59
Web Application ... 60
Pipeline ... 60
Preprocessing ... 61
MIC Computation ... 62
Graph Creation ... 63
Neo4j Database Design ... 64
Graph Analysis ... 65
Additional Post-processing ... 67
Chapter 8 - Results ... 69
Web Application ... 69
Network Visualizations ... 70
Node Counts ... 75
Edge Counts ... 80
MIC Distribution ... 85
Edge Densities ... 91
Degree Distribution ... 92
Clustering Coefficient Distributions ... 99
Top 25 Nodes ... 104
Top 25 Edges ... 114
Entropy ... 119
Solvent Accessibility ... 122
Residue Cooccurence Counts ... 123
Protein Correlation Graphs ... 124
Evolution of Human H1N1 Network ... 130
Evolution of Human H3N2 Network ... 132
Chapter 9 – Conclusions ... 135
Network Analysis of Correlated Mutations in IAV can provide novel insights ... 135
Correlated Mutation Networks in IAV are sub-type and host specific ... 136
Correlated Mutation Networks can improve our understanding of Influenza evolution ... 136
Observations from Correlated Mutation Networks in Influenza (@ 0.5 threshold) ... 136

There is a statistically significant correlation between entropy and MIC correlations 137

There is no statistically significant correlation between solvent accessibility and MIC correlations 137

Computational pipeline and software can be applied to other systems 137

Normalized 3-gram counts are reliable features for automated classification of IAV sequences 138

Chapter 10 - Future Work 139

Chapter 11 - Appendix 141

Intra-Protein Correlated Mutations in IAV 141

Source Code and Web Application 142

References 143

LIST OF TABLES

Table Page
Table 1 - Datasets used for Classification.. 26
Table 2 - Important features in M2 (based on AD resistance vs AD sensitivity)............. 33
Table 3 - Important features in NA (based on OS resistance vs OS sensitivity).............. 33
Table 4 – Primary Datasets used for Network Analysis of Correlated Mutations............ 52
Table 5 - Computational Tools and Libraries.. 55
Table 6 - Neo4j labels and properties 65
Table 7 - Networks @ 0.5.. 91
Table 8 – Top 25 residues for Swine H1N1 Dataset 106
Table 9 - Top 25 nodes for Human H1N1 Dataset 107
Table 10 - Top 25 nodes for Human H3N2 Dataset 108
Table 11 - Top 25 residues for Swine H3N2 Dataset 110
Table 12 - Top 25 residues for Avian H5 Dataset 111
Table 13 - Top 25 residues for H7N9 Dataset 112
Table 14 - Top 25 residues for Human-All Dataset... 113
Table 15 - Top 25 edges for Swine H1N1 dataset... 115
Table 16 - Top 25 edges for Human H1N1 Dataset 115
Table 17 - Top 25 edges for Swine H3N2 Dataset 116
Table 18 - Top 25 edges for Human H3N2 Dataset 117
Table 19 - Top 25 edges for Avian H5 dataset 118
Table 20 - Top 25 edges for H7N9 dataset.. 118
Table 21 - Residue Cooccurence Counts 124
Table 22 - Datasets for studying Human H1N1 evolution 130
Table 23 - Datasets for studying Human H3N2 evolution 132

LIST OF FIGURES

Figure Page

Figure 1 – Structure of Influenza [20] ... 14

Figure 2 - Replication Cycle of Influenza ... 17

Figure 3 - Classification accuracy for adamantane resistant vs adamantane sensitive sequences of HA and M2 using four different classification algorithms ... 31

Figure 4 - Classification accuracy for oseltamavir resistant vs oseltamavir sensitive sequences of HA and NA using four different classification algorithms ... 31

Figure 5 - Classification accuracy for adamantane-resistant, oseltamavir-sensitive vs adamantane-sensitive, oseltamavir-resistant sequences of HA, NA, M1& M2 using four different classification algorithms ... 31

Figure 6 - weblogo for a 10-residue segment around the most significant 3-gram (227/228/229) in HA ... 34

Figure 7 - weblogo for a 11-residue segment around the most significant 3-gram (274/275/276) in NA ... 34

Figure 8 - weblogo for an 8-residue segment around the most significant 3-gram in M2 34

Figure 9 - ROC plots based on classification of AD resistant and AD sensitive sequences of HA, values in parenthesis represent area under curve (AUC) ... 35

Figure 10 - ROC plots based on classification of OS resistant and OS sensitive sequences of NA, values in parenthesis represent area under curve (AUC) ... 36

Figure 11 - Classification accuracy for H1N1 vs H3N2 classification ... 36

Figure 12 - Classification accuracy based on geographic location ... 37

Figure 13 - Classification accuracy for 2013-14 vs 2014-15 sequences ... 38

Figure 14 - Classification accuracy for protein sequences based on pH1N1 similarity ... 38

Figure 15 - Classification accuracy for adamantane-resistant, oseltamavir-sensitive vs adamantane-sensitive, oseltamavir-resistant DNA sequences of HA, NA, M1& M2 using four different classification algorithms ... 39

Figure 16 - Classification accuracy based on year of infection for HA, NA, M1& M2 DNA sequences using four different classification algorithms ... 40

Figure 17 - MIC inter-residue correlation scores between NA, HA and NA, M1 proteins. x and y axis represent residue numbers within that protein ... 41

Figure 18 - MIC inter-residue correlation scores between NA, H1 and NA, M1 proteins. x and y axis represent residue numbers within that protein ... 42

Figure 19 - Search options in the IRD ... 50

Figure 20 - Search Results in the IRD ... 51

Figure 21 - Download Results in the IRD ... 52

Figure 22 – Overall Pipeline for Network Analysis of Correlated Mutations. ... 61

Figure 23 – Preprocessing step where strains of IAV are downloaded, de-duplicated and split into individual protein-specific sequence files to enable downstream processing ... 62

Figure 24 - MIC Computation pipeline where pairwise residue MIC computation is performed between residues in two proteins ... 63

Figure 25 - Graph Creation pipeline where graphs are created based on inter-residue correlated mutations 64
Figure 26 - Graph Analysis pipeline for analysis of correlated mutation graphs 67
Figure 27 - Additional post processing steps that we conducted to provide additional context to our results 68
Figure 28 - Visualization of human H3N2 correlated mutation network at 0.5 threshold 71
Figure 29 - Visualization of swine H3N2 network at 0.5 threshold 72
Figure 30 - Visualization of human H1N1 network at 0.5 threshold........ 73
Figure 31 - Visualization of swine H1N1 network at 0.5 threshold 74
Figure 32 - Visualization of avian H5 network at 0.5 threshold........ 75
Figure 33 - Node counts for H1N1 datasets........ 77
Figure 34 - Node counts for H3N2 datasets........ 78
Figure 35 - Node counts for AVIAN H5 dataset 78
Figure 36 - Node counts for H7N9 dataset 79
Figure 37 - Node counts for HUMAN_ALL dataset 80
Figure 38 - Edge counts for H1N1 datasets 82
Figure 39 - Edge counts for H3N2 datasets 82
Figure 40 - Edge counts for AVIAN H5 dataset........ 83
Figure 41 - Edge counts for H7N9 dataset........ 84
Figure 42 - Edge counts for HUMAN ALL dataset 85
Figure 43 - MIC distribution for SWINE H1N1 dataset........ 86
Figure 44 - MIC distribution for HUMAN H1N1 Dataset 87
Figure 45 - MIC distribution for SWINE H3N2 dataset........ 87
Figure 46 - MIC distribution for HUMAN H3N2 Dataset 88
Figure 47 - MIC distribution for AVIAN H5 dataset 89
Figure 48 - MIC distribution for H7N9 dataset 90
Figure 49 - MIC distribution for HUMAN_ALL dataset 91
Figure 50 - Degree histogram for Swine H1N1 dataset........ 93
Figure 51 - Degree distribution for Human H1N1 dataset 94
Figure 52 - Degree histogram for Swine H3N2 dataset........ 95
Figure 53 - Degree distribution for Human H3N2 dataset 96
Figure 54 - Degree distribution for Avian H5 dataset 97
Figure 55 - Degree distribution for H7N9 dataset 98
Figure 56 - Degree distribution for Human All dataset 99
Figure 57 - Average Clustering Coefficient........ 100
Figure 58 - Clustering coefficient histogram for Swine H1N1 dataset 101
Figure 59 - Clustering coefficient histogram for Human H1N1 dataset........ 101
Figure 60 - Clustering coefficient histogram for Swine H3N2 dataset 102
Figure 61 - Clustering coefficient histogram for Human H3N2 Dataset........ 103
Figure 62 - Clustering coefficient histogram for Avian H5........ 103
Figure 63 - Clustering coefficient histogram for H7N9 dataset 104
Figure 64 - Clustering coefficient distribution for Human-All dataset........ 104
Figure 65 - Average Entropies........ 120
Figure 66 - Average Entropies for In-Network Residues 121

Figure 67 - Average Entropies for Out-of-Network Residues 122
Figure 68 - Average solvent accessibility for in-network and out-of-network residues. 123
Figure 69 - Protein Correlation Network for Swine H1N1 Dataset 126
Figure 70 - Protein Correlation Network for Human H1N1 Dataset 126
Figure 71 - Protein Correlation Network for Swine H3N2 Dataset 127
Figure 72 - Protein Correlation Network for Human H3N2 Dataset 128
Figure 73 - Protein Correlation Network for Avian H5 Dataset 128
Figure 74 - Protein Correlation Network for H7N9 Dataset 129
Figure 75 - Protein Correlation Network for Human_All Dataset 129
Figure 76 - Node counts for Human H1N1 Datasets 131
Figure 77 - Edge counts for Human H1N1 Datasets 132
Figure 78 - Node counts for Human H3N2 Datasets 133
Figure 79 - Edge counts for Human H3N2 Datasets 134
Figure 80 - Node counts in intra-protein correlated mutation graphs with MIC threshold set to 0.4 141
Figure 81 - Edge counts in intra-protein correlated mutation graphs with MIC threshold set to 0.4 142

LIST OF ABBREVIATIONS

IAV	Influenza A Virus
IRD	Influenza Research Database
HA	Hemagglutinin
NA	Neuraminidase
M1	Matrix1
M2	Matrix2
NP	Nucleocapsid Protein
PA	Polymerase Acidic Protein
NS1 and NS2	Non Structural Proteins
CNA	Correlation based Network Analysis
MIC	Maximal Information Coefficient

ABSTRACT

NETWORK ANALYSIS OF CORRELATED MUTATIONS IN INFLUENZA

Uday Yallapragada, Ph.D.

George Mason University, 2017

Dissertation Director: Dr. Iosif Vaisman

Influenza A Virus (IAV) is remarkably adept at surviving in human populations. IAV thrives even among populations with wide spread access to vaccines and anti-viral drugs, and continues to be a major cause of morbidity and mortality. Correlated mutations are an important factor in IAV's evolution and are critical for host adaptation and pathogenicity. Large sets of publicly available sequences of IAV combined with its rapid and complex evolutionary dynamics present interesting opportunities and unique challenges to analyze correlated mutations in influenza proteomes. In this work, we performed a comprehensive analysis of correlated mutations in IAV using a network theory approach where residues in each protein act as nodes in the graph and edges in the graph are created based on inter-residue correlated mutations. Our approach used 'maximal information coefficient' (MIC) to compute correlations between residues and we created edges between nodes if MIC exceeds a threshold. We created a modular and robust pipeline and applied it to multiple datasets belonging to H1N1, H3N2, H5 and H7N9 subtypes. We

studied structural dynamics of IAV sub-systems based on topological properties of their networks resulting in several important conclusions. We identified nodes with highest degree along with edges and triplets with strongest weight for each network. To contextualize our results, we performed entropy analysis to gain a global view of sequence variation and computed solvent accessibility profiles to identify statistical differences in correlation profiles between surface and buried residues. We computed residue cooccurrence counts to understand the internal mechanics behind MIC. Additionally, we applied our pipeline to gradually increasing datasets of human H1N1 and human H3N2 over the past 10 years and elucidated their evolutionary patterns. As part of our overall pipeline, we took specific measures to eliminate phylogenetic and stochastic background noise. We created a web application to allow users to comprehend results of our analysis and to search for correlated mutations.

CHAPTER 1 - INTRODUCTION

Dynamic Nature of Influenza A Virus

Influenza A Virus (IAV) is remarkably adept at surviving in the human population over a long-time scale. The human IAV continues to thrive even among populations with widespread access to vaccines and continues to be a major cause of morbidity and mortality [1]. According to World Health Organization, influenza occurs globally with an annual attack rate estimated at 5%-10% in adults and 20%-30% in children and it is estimated to result in about 3 to 5 million cases of severe illness and about 250000 to 500000 deaths annually [2]. The virus mutates from year to year making the existing vaccines ineffective on a permanent basis and requiring that new strains be chosen for a new vaccine. Less frequent major changes known as antigenic shift create new strains against which human population has little protective immunity thereby causing pandemics. These high evolutionary rates are also responsible for gradually increasing resistance exhibited by the virus to existing antiviral drugs.

Correlated Mutations

‘Correlated Mutations’ in the structural context were first introduced in 1987 [3]. In 1994 Shindyalov et al. [4] hypothesized that three-dimensional contacts can be predicted by analyzing correlated mutations. This study was immediately followed by [5] where the term “correlated mutation” with a protein context was formally defined as ‘tendency of positions in proteins to mutate coordinately'. This tendency is measured by analyzing the correlation between changes in pairs of positions in multiple sequence alignment. Sequence correlation/covariance analysis is an area that has gained significant traction over the last two decades and is widely used for identifying correlated sites in proteins. In this context, it is important to note that correlated mutations between residues in proteins have initially been linked primarily to probable physical contact in three-dimensional space but more recent studies have demonstrated that coevolution of amino acids may originate not only from structural contacts but also from a much broader range of biological reasons. More specifically, studies in this field have suggested that

(a) correlated mutations may also occur due to reasons related to protein function [6]

(b) coevolution between amino acid residues is necessary and sufficient to specify sequences that fold into native structures [7]

(c) and residues highly correlated with others are indeed more likely to be associated with disease [8].

Correlated Mutation Analysis in IAV Sequences

Changes in the genetic makeup of IAV are primarily attributed to antigenic drift that results in accumulation of mutations both within the genes that code for anti-body binding sites and sites that are not directly targeted by anti-bodies. Proteins in IAV are a natural fit for comprehensive correlation analysis because of the large sets of publicly available sequences available, and this observed (fast) rate of mutations in these proteins.

Majority of the correlation studies on influenza up to date have focused on identifying pairwise mutations within the transmembrane proteins HA and NA using specific statistical and/or machine learning techniques. Moreover, some of these studies have been conducted at a time when there were only limited sequences available. A comprehensive analysis of pairwise and higher-order correlations in mutations using existing data sets can provide a more holistic picture of potential functional implications.

Network Analysis

While there is a significant body of literature on the analysis of correlated mutations in proteins in general, there was very little work done up to date to comprehensively analyze correlations in IAV strains. Majority of the work performed on IAV strains either focused on the two surface proteins (HA and NA) or was based on small datasets. After performing extensive review of available literature on correlated mutations in general and more specifically on IAV strains (Chapter 4 - Literature Review) it became apparent that the large number of curated IAV strains available in public databases presents a unique opportunity to conduct a thorough analysis of mutational landscape. We conceived a novel approach based on principles in graph theory and network analysis where residues in each protein act as nodes in the graph and edges in the graph are created based on inter-residue correlated mutations. Our approach used ‘maximal information coefficient’ (MIC) to compute correlations between residues and we created edges between nodes if MIC exceeds a certain threshold value.

Significance

The broader implications of creating a comprehensive correlation profile based on network analysis for mutations in IAV can be broken into three broad categories.

We believe that our work can have both direct and indirect impact on future work performed in these areas.

Antiviral Drugs

Analogous to the treatments used against HIV-1 infection, it has been hypothesized [9] that a combination of potent influenza drugs, each targeting a different viral entity or with different modes of inhibition, would be expected to be more effective in treating virulent and pandemic influenza viruses. A recent study employed a network based approach for determining conserved amino acid sequences within a protein system to more effectively identify epitopes on viral proteins (for the design/identification of broadly-neutralizing monoclonal antibodies or specific immunogens for anti-viral drugs and vaccine development) [10]. Given the fact that viruses universally develop mutations that allow escape from neutralization, the above study suggests that protein function may not be dependent on observed conservations at point locations. Potential links between conserved sites within different proteins can be exploited to create a more potent combination of drugs and for effective epitope identification. Correlations between mutations in proteins that are co-involved in a specific step or sub-process within the influenza life cycle may provide new insights into potential interactions that can be useful for drug design.

Vaccines

Current influenza vaccines induce neutralizing antibodies against the viral membrane surface proteins hemagglutinin (HA) and neuraminidase (NA). Due to antigenic shift and drift of HA and NA genes, neutralizing antibodies elicited by influenza vaccines lack cross-reactivity against non-matching influenza strains. While seasonal adjustments to the vaccine strains are made to cope with this problem, it is not as convenient and fast as a potential cross-protective influenza vaccine. Thus, the identification of alternative correlates of protection (CoPs) against influenza is an important step toward the development of cross-reactive influenza vaccines. Developing a universal influenza vaccine based on a more conserved part of the influenza virus which is not affected by antigenic change or that is consistent across all strains remains the ultimate goal to afford cross-protection [11]–[14].

Epidemiology

Several fundamental questions regarding the epidemiology of influenza remain unanswered [15] and there is a continued interest in having a more comprehensive understanding of the impact of epistasis on broader epidemiological patterns of this virus. Having a broader understanding of the overall mutation profile of IAV based on different datasets can provide insights into these patterns.

Objectives

The primary objective of this work is to provide novel insights into the functional dynamics of IAV system based on network analysis of correlated mutations. We perform this analysis on multiple datasets and try to examine the differences. Based on our analysis, we attempt to address the following problems.

General properties of the networks of correlated mutations

1. What is the overall structure of the network of correlated mutations?
2. What is the behavior of the network for various threshold values?
3. Does a correlated mutation graph follow the power-law model?
4. What is the overall degree and clustering distribution? What is the diameter? Is this a “small world” network?
5. What is the relationship between global characteristics of the graph (degree distribution, edge density) and entropy profiles of the residues? Can we hypothesize that residues with zero or very low entropies will be out-of-network while in-network residues will always have higher entropies?
6. Are there differences in residue cooccurrence counts based on correlation scores?

Subtype dependence of the network properties

- Are there significant differences between networks of correlated mutations depending on the virus subtype? E.g., how does the human H1N1 mutation graph differ from swine H1N1 mutation graph? Do the significant residues in human H3N2 mutation graph differ substantially from the significant residues in swine H3N2 mutation graph?
- What are the characteristics of a mutation graph constructed from strains belonging to multiple subtypes? Does this graph differ significantly from graphs created from a single subtype?

Protein structure and function implications of network features

1. Is there a significant overlap between the known functionally important residues in influenza and the nodes with highest degree?
2. Is there a significant difference in solvent accessibility profile between in-network residues and out-of-network residues?
3. Are there more edges between the two surface proteins (HA and NA) and between HA, NA and other proteins compared to edges between non-surface proteins?
4. How does the Human H1N1 network evolve over multiple flu seasons?

Computational contributions

From a computational perspective, this dissertation offers the following contributions.

1. Web application – We created a user-friendly web application to allow users to examine our results and visualizations and to perform basic search for correlated sites of a specific residue.
2. Source code & datasets – All datasets used for this analysis and python code developed are made publicly available.
3. This effort endeavored to create a robust computational pipeline that can potentially be applied to perform network analysis of other viral systems in the future. Artifacts created as part of this initiative should have applicability in the analysis of similar viral systems like Ebola, Chikungunya, West Nile and Enterovirus.

Dissertation Structure

After elucidating the opportunity, objectives and significance of our work in this Chapter, we provide required biological background of influenza virus in Chapter 2. We start with basic information on influenza classification and follow it up with a broad overview of structural details and various steps involved in the life cycle of influenza. Chapter 2 also provides additional background information on existing vaccines, anti-viral drugs and epidemiology. Given the extensive use of concepts from graph theory in our dissertation, we provide relevant background information on concepts and principles from graph theory and network analysis in Chapter 2. We also include details of MIC in this chapter since we use MIC to calculate correlation coefficient throughout our work. In Chapter 3, we provide details of work that we conducted to automatically classify flu sequences using n-grams. The results of this work on classification have motivated us to perform a comprehensive analysis of correlated mutations in IAV sequences. In Chapter 4, we provide a broad overview of literature available in the areas of correlated mutations and network analysis. Chapter 5 covers details of datasets and Chapter 6 provides a listing of computational tools and libraries that we used in this work. Chapter 7 elucidates the methodology that we used to identify and analyze correlated mutations. We begin this chapter by providing a pipeline of the end-to-end flow and we explain each of the steps in this process. We cover details of the three important components of our computational pipeline - pre-processing, graph creation and graph analysis in this chapter. Chapter 8 focusses on summarizing results of our work. We attempt to answer questions listed as part of our objectives. Our analysis covers the macro properties of the system based on network analysis as well as delves into identifying and studying significant nodes and edges. We compare the properties of different IAV sub-systems based on network analysis. In Chapter 8, we also explain the functionality of a web application called NACMI (**N**etwork **A**nalysis of **C**orrelated **M**utations in **I**nfluenza) that we created to allow users to perform analysis and perform basic search for correlated mutations in influenza. Chapter 9 offers our conclusions and Chapter 10 includes suggestions for future work.

CHAPTER 2 - ADDITIONAL BACKGROUND

In Chapter 1, we have stated that our work will focus on performing network analysis of correlated mutations in influenza strains. Before we delve into the details of datasets, methodology and results of our work, we provide some necessary background information on influenza, Network Analysis and Maximal Information Coefficient (MIC) in this chapter. We start with an overview of the classification, structure and life cycle of the Influenza virus. Next, we describe some key concepts from graph theory and network analysis. We conclude this chapter with an introduction to correlations and MIC.

Influenza

Influenza (flu) is a contagious respiratory illness caused by Influenza viruses. It can cause mild to severe illness. Serious outcomes of flu infection can result in hospitalization or death. Certain categories of people, such as older people, young children and people with certain health conditions are at high risk for serious flu complications.

Classification of Influenza Virus

Influenza virus types A, B and C belong to the family of Orthomyxoviridae and have negative sense, single-stranded, segmented RNA. The most prominent difference between the three types of influenza virus is the host range. While influenza viruses of types B and C are predominantly human pathogens that have sporadically been isolated from seals and pigs respectively, IAVs have been isolated from many animal species, including humans, pigs, horses and a wide range of birds. Influenza virus B mutates at a rate 2–3 times slower than type A and consequently is less genetically diverse, with only one influenza B serotype. Because of this lack of antigenic diversity, a degree of immunity to influenza B is usually acquired at an early age. However, influenza B mutates enough that lasting immunity is not possible. This reduced rate of antigenic change, combined with its limited host range (inhibiting cross species antigenic shift), ensures that pandemics of influenza B do not occur. Influenza C is less common than A and B types and usually only cause's mild disease in children. *The type A viruses are the most virulent human pathogens among the three influenza types and cause the most severe disease. Our work will focus only on correlated mutations in IAV strains.* Wild aquatic birds are the natural hosts for a large variety of influenza A strains. Occasionally, these viruses are transmitted to other species and may then cause devastating outbreaks in domestic poultry or give rise to human influenza pandemics. The main antigenic determinants of influenza A and B viruses are the Hemagglutinin (HA or H) and Neuraminidase (NA or N) transmembrane glycoproteins. Based on the antigenicity of these glycoproteins, Influenza A viruses are further subdivided into sixteen H (H1-H16) and nine N (N1-N9) subtypes. Example subtypes of Influenza A include H5N1 (bird flu) and H1N1 (swine flu) [16].

Nomenclature of Influenza Strains

The strain designation for influenza virus types A, B, and C contains the following information:

1. A description of the antigenic type of the virus based on the antigenic specificity of the NP antigen (type A, B, or C). Since 1971, a further type-specific internal antigen of the influenza A and B viruses, the matrix (M) protein, has been described (19).

Typing of Influenza, A and B viruses based on the M protein is consistent with the results obtained with NP antigen (20).

2. The host of origin. This is not indicated for strains isolated from human sources but is indicated for all strains isolated from non-human hosts, e.g., swine, horse (equine), chicken, and turkey.
3. Geographical origin.
4. Strain number.
5. Year of isolation.
6. For influenza A viruses, the antigenic description, in parentheses, follows the strain designation and includes the following information.
    a. An index describing the antigenic character of the Hemagglutinin, i.e., H1, H2, H3, H4, etc. The numbering of subtypes is a simple sequential system which applies uniformly to influenza viruses from all sources.
    b. An index describing the antigenic character of the Neuraminidase, i.e., N1, N2, N3, N4, etc. As with the H antigen subtype, this is a simple sequential numbering system applied uniformly to all Influenza A viruses.
7. Examples
    a. A/swine/Virginia/01359/2006 (H1N1),
    b. A/green-winged teal/Minnesota/Sg-00820/2008 (H4N5) [17]

Proteins in Influenza

Influenza A and B viruses are enveloped viruses with eight RNA segments that encode for the following 11 proteins: glycoproteins (hemagglutinin (HA) and neuraminidase (NA)), matrix 1 (M1), matrix 2 (M2), nucleoprotein (NP), non-structural protein 1 (NSP1), non-structural protein 2 (NS2; also known as nuclear export protein, NEP), polymerase acidic protein (PA), polymerase basic protein 1 (PB1), polymerase basic protein 2 (PB2) and polymerase basic protein–F2 (PB1-F2) [18], [19].

Influenza C virus contains only 7 RNA segments that encode for 10 proteins with only one glycoprotein (HEF, hemagglutinin-esterase-fusion) that has the functionality of both HA and NA [18], [19]

Structure of Influenza

The structure of Influenza A virus is depicted in Figure 1. The Influenza virion is roughly spherical. It is an enveloped virus with an outer lipid membrane with 'spikes' consisting of two surface glycoproteins – HA and NA. The HA binds the virus to sialic acid receptors on the host cell surface. The NA protein facilitates the release of virions to infect other cells by removing sialic acid residues from the viral HA during entry and release from cells [18], [19].

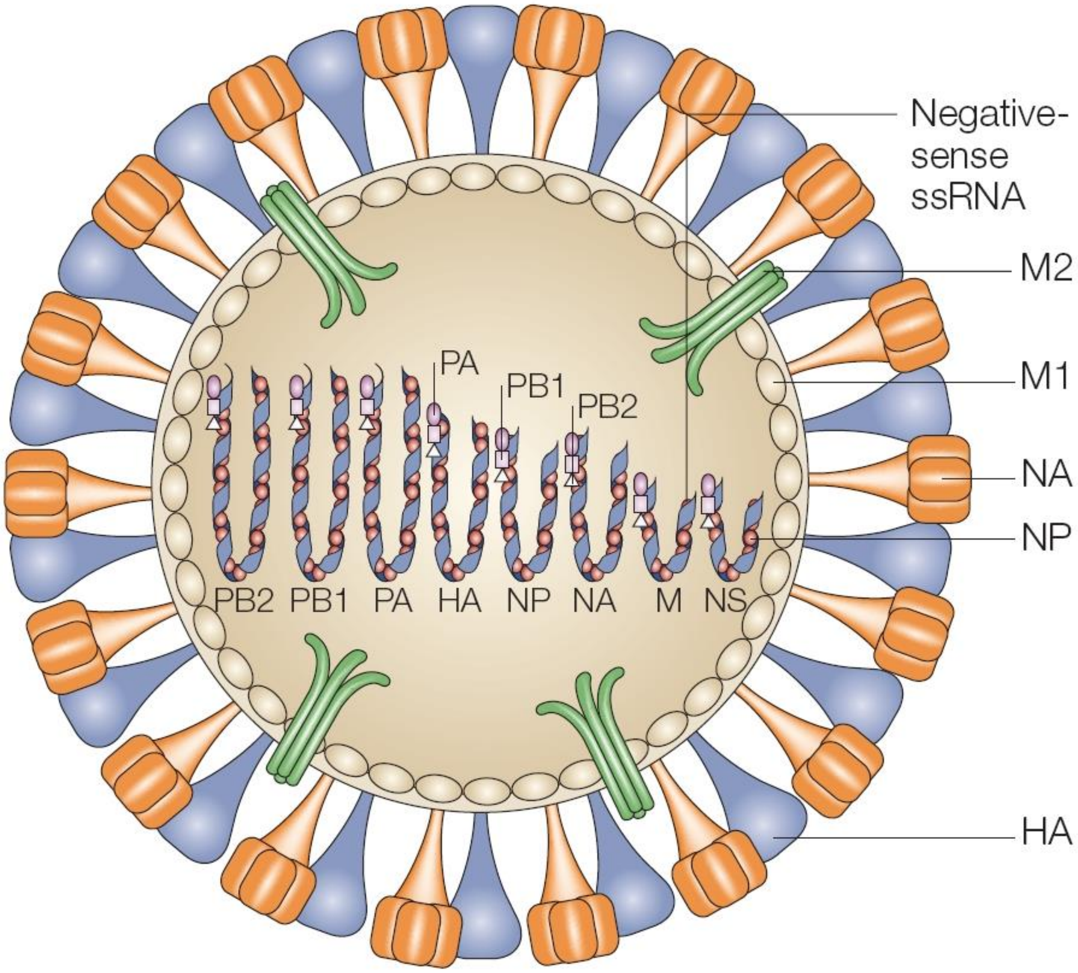


Figure 1 – Structure of Influenza [20]

Beneath the lipid membrane is a viral protein called M1, or matrix protein. This protein, which forms a shell, gives strength and rigidity to the lipid envelope. The IAV envelope contains a small number of a membrane protein, M2 which forms a tetramer with ion channel activity. M2 is involved in the infection process by modulating the pH within virions, weakening the interaction between the viral ribonucleoproteins (RNPs) and the M1 protein. Within the interior of the virion are the viral RNAs – 8 of them for Influenza A viruses. These are the genetic material of the virus; they code for one or two proteins. Each

RNA segment consists of RNA joined with several proteins shown in the diagram: PB1, PB2, PA and NP. These RNA segments are the genes of Influenza virus.

Life Cycle of Influenza

Influenza viruses are usually transmitted via air droplets, and subsequently contaminate the mucosa of the respiratory tract. They can penetrate the mucin layer of the outer surface of the respiratory tract, entering respiratory epithelial cells, as well as other cell types. Immuno-histological pictures show that foci of virus-producing cells are clustered in the mucous layer of the respiratory tract, in the gut and even in endothelial layers, myocardium and brain. Within nasal secretions, millions of virus particles per ml are shed, so that a 0.1 µl aerosol particle contains more than 100 virus particles. A single HID (human infectious dose) of influenza virus might be between 100 and 1,000 particles. At least during the early course of influenza infection, the virus can be found also in the blood and in other body fluids. Replication is very quick: after only 6 hours the first influenza viruses are shed from infected cells

Influenza infection and replication is a multi-step process depicted in Figure 2. First, the virus must bind to and enter the cell, then deliver its genome to a site where it can produce new copies of viral proteins and RNA, assemble these components into new viral particles, and, last, exit the host cell.

Influenza viruses bind through hemagglutinin onto sialic acid sugars on the surfaces of epithelial cells, typically in the nose, throat, and lungs of mammals, and intestines of birds (Stage 1 in Figure 2). After the hemagglutinin is cleaved by a protease, the cell imports the virus by endocytosis.

Once inside the cell, the acidic conditions in the endosome cause two events to happen: First, part of the hemagglutinin protein fuses the viral envelope with the vacuole's membrane, then the M2 ion channel allows protons to move through the viral envelope and acidify the core of the virus, which causes the core to disassemble and release the viral RNA and core proteins. The viral RNA (vRNA) molecules, accessory proteins and RNA-dependent RNA polymerase are then released into the cytoplasm (Stage 2).

These core proteins and vRNA form a complex that is transported into the cell nucleus, where the RNA-dependent RNA polymerase begins transcribing complementary positive-sense vRNA (Steps 3a and b). The vRNA either is exported into the cytoplasm and translated (step 4) or remains in the nucleus. Newly synthesized viral proteins are either secreted through the Golgi apparatus onto the cell surface (in the case of neuraminidase and hemagglutinin, step 5b) or transported back into the nucleus to bind vRNA and form new viral genome particles (step 5a). Other viral proteins have multiple actions in the host cell, including degrading cellular mRNA and using the released nucleotides for vRNA synthesis and inhibiting translation of host-cell mRNAs.

Negative-sense vRNAs that form the genomes of future viruses, RNA-dependent RNA polymerase, and other viral proteins are assembled into a virion. Hemagglutinin and neuraminidase molecules cluster into a bulge in the cell membrane. The vRNA and viral core proteins leave the nucleus and enter this membrane protrusion (step 6). The mature virus buds off from the cell in a sphere of host phospholipid membrane, acquiring hemagglutinin and neuraminidase with this membrane coat (step 7). As before, the viruses adhere to the cell through hemagglutinin; the mature viruses detach once their neuraminidase has cleaved sialic acid residues from the host cell. After the release of new influenza viruses, the host cell dies [19], [21].

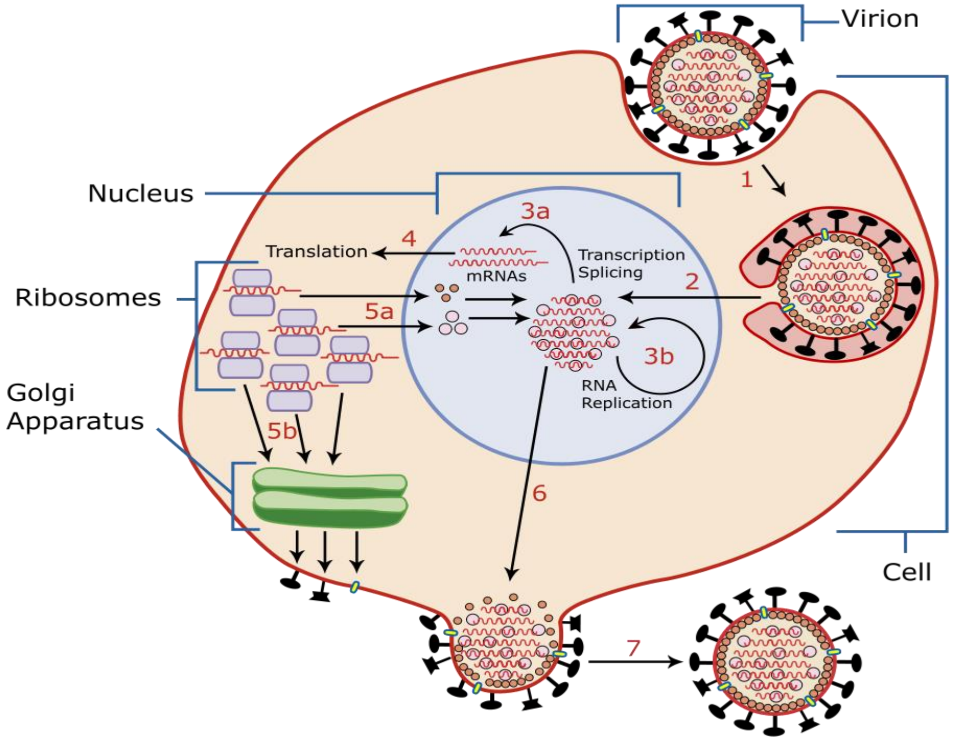


Figure 2 - Replication Cycle of Influenza

Network Analysis

A network is any collection of objects in which some pairs of these objects are connected by links [22]. Given the flexibility of this definition, it is easy to find networks in many domains and the concept of networks is finding increasing applicability to analyze and predict the structure and dynamics of complex systems. We will discuss several applications of network analysis to understand biological systems and more specifically protein systems in *Chapter 4 - Literature Review*. In this section, we introduce concepts from “graph theory” since the study of network structure relies on principles from graph theory.

A graph is a way of specifying relationships among a collection of items. A graph consists of a set of objects called nodes with certain pairs of these objects connected by links called edges. Two nodes are neighbors if they are connected by an edge. An undirected graph has edges that can be traversed in both directions while a directed graph has edges that can only be traversed in one direction ([23], [24]). In a weighted graph, an edge has a number weight associated with it to denote the strength of association between the nodes. In an unweighted graph, all edges are equivalent (or have the same weight). A connected graph is one in which it is possible to go between any pair of nodes via a path through a series of edges and other nodes. A completely connected graph is one where every node is directly connected by an edge to every other node.

After a network is created with appropriate nodes and edges, a variety of useful measures can be calculated to capture the structural topology of the network. In this work, we focus primarily on two measures called *degree* and *clustering coefficient*.

Degree (or degree centrality) of a node is the simplest measure in a network and is the number of edges connected to that node. In the context of correlated mutation networks, degree of a node is a useful measure since it is reasonable to assume that a node with high degree can act as a hub and has more influence in the network compared to a node with low degree. More importantly, degree distribution of nodes in a correlated mutation network can provide a good overview of the overall mutation profile of residues in proteins.

Clustering coefficient (or local clustering coefficient) of a node in a network is defined as the ratio of the number of pairs of neighbors of that node that are connected and the total number of possible pairs of neighbors. In the context of a correlated mutation network, a node with high clustering coefficient signifies that this node is part of a highly connected sub-graph where there is strong covariance between multiple residues. Nodes with high clustering coefficient generally tend to have a low degree and vice-versa.

We have reviewed several other measures of networks to understand topologies of correlated mutation graphs (including page rank, betweenness centrality, closeness centrality and eigen value centrality) and have concluded that degree and clustering coefficients are more appropriate.

Correlation Measures

Our approach to network analysis of correlated mutations relies on computation of a correlation measure to assign a weighted edge between nodes. The choice of a correlation measure that can accurately capture complex linear and non-linear relationships and quantify it appropriately is important. After careful analysis of several available correlation measures [25], we have decided to use a relatively novel technique called “Maximal Information Coefficient” for this work. Linear dependence measures such as Pearson correlation or monotonic dependence measures such as Spearman’s do not capture complex relationships in biological systems [26]. While Mutual Information (MI) can identify non-linear relationships in data, MI has been sensitive to bin size and number of bins in some cases and more importantly MI has an unsatisfying (0 to infinity) range that limits its applicability when being used broadly for multiple datasets. The MIC [27] solves both of these issues and is being increasingly used on biological datasets ([26], [28]).

Maximal Information Coefficient

The MIC is a tool for finding the strongest pairwise relationships in a data set with many variables. MIC is useful because it gives similar scores to equally noisy relationships of different types. This property, called equitability, is important for analyzing high-dimensional data sets. MIC is based on the idea that if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship. Thus, to calculate the MIC of a set of two-variable data, we explore all grids up to a maximal grid resolution, dependent on the sample size, computing for every pair of integers (x, y) the largest possible mutual information achievable by any x-by-y grid applied to the data. We then normalize these mutual information values to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1. We define the characteristic matrix $M = (m_{x, y})$, where $m_{x, y}$ is the highest normalized mutual information achieved by any x-by-y grid, and the statistic MIC to be the maximum value in M.

CHAPTER 3 – CLASSIFICATION OF INFLUENZA SEQUENCES

Introduction

In this chapter, we provide details of related work that we conducted to automatically classify flu sequences using n-grams. The results of this work on classification have motivated us to perform a comprehensive analysis of correlated mutations in IAV sequences.

Nucleotide sequence variation among genomes of highly variable, fast mutating pandemic viruses presents a challenge for systemic classification. Fast, accurate and reliable classification of IAV sequences can help determine routes of disease transmission, provide therapeutic and clinical decision support, facilitate monitoring for new variants, improve our understanding of relationships between IAV diversity and immune response and can simplify analysis of treatment-resistant mutations. The gradual increase in number of sequences in IRD [29] presents a unique opportunity to apply state-of-the-art machine learning algorithms to accurately classify IAV sequences based on specific characteristics. To date, most of the work done in this area focused on classifying hemagglutinin (HA) and neuraminidase (NA) sequences in IAV with respect to clade and subtype. In this effort, we attempted to classify nucleotide sequences of Matrix proteins (M1 and M2), Polymerase basic protein 1 (PB1), and Nucleoprotein (NP) along with HA and NA based on multiple characteristics – drug resistance, subtype, year of infection, geography and pH1N1 (similarity to 2009 pandemic H1N1). We feed normalized n-gram frequency counts of IAV protein and nucleotide sequences as feature vectors to four different classification algorithms - Random Forest (RF), Support vector machine (SVM), Gaussian naïve bayes (GNB) and K-Nearest Neighbors (KNN). We apply this approach on several different data sets to perform binary classification. To our knowledge, this is the most comprehensive effort up to classify IAV sequences.

Background

Influenza A is among the most extensively studied viruses because of its importance as a human pathogen. With a large, public database of genetic sequences, Influenza virus offers to be an appropriate system for studying antigenic and genetic drift in general. The evolution of Influenza viruses is characterized by frequent re-assortment events within subtypes as well as high rates of amino-acid substitutions in several proteins including the viral surface proteins HA and NA. Such high evolutionary rates reflect both poor fidelity of the viral proteins, and strong selection pressures to evade the human immunity and present unique challenges in terms of accurately classifying sequences. ClassyFlu is a recently developed tool for the classification of IAV sequences of the HA and NA gene into subtypes and phylogenetic clades using discriminatively trained profile hidden Markov models [30]. The IRD proposed ‘‘Highly Pathogenic H5N1 Clade Classification Tool’’ (IRDCT) as a free web service for the classification of IAV sequences into phylogenetic clades. IRDCT is based on phylogeny but keeps the tree of already classified sequences fixed.

In the absence of an effective and widely available Influenza vaccination, Anti-influenza drugs serve as valuable “second line of defense” to treat Influenza. Two classes of anti-influenza drugs - Adamantanes (amantadine and rimantadine) and Neuraminidase inhibitors (oseltamivir and zanamivir) are currently approved by FDA for treatment of influenza [31]. Adamantanes act by blocking the ion channel formed by M2 protein and inhibiting early stages of virus replication while Neuraminidase inhibitors attenuate patient infection by binding to active site on viral Neuraminidase (NA) and blocking its ability to infect additional cells. The main drawback of utilizing Adamantanes is that drug-resistant variants develop rapidly. A study conducted in 2007 reported a 15.5% Adamantane resistance among H1N1 viruses and a much higher percentage for H3N2 viruses isolated worldwide in 2005-2006 [32]. A more recent study [33] found that 45.2% of Influenza A viruses circulating globally were resistant to Adamantanes and the vast majority of these resistant virus bear S31N mutation in M2 protein sequence. Few cases of resistance to oseltamivir in strains of the pandemic Influenza A 2009 (H1N1) virus have been reported worldwide [34]. In [35], Nguyen et al. reported an Influenza A 2009 H1N1 virus strain with laboratory evidence for Oseltamavir & Zanamivir resistance in a 14-year-old girl with 2 NA mutations (H275Y and I223R). Reports of resistance to NA inhibitors have also been reported more recently in H7N9 IAV strains [36], [37].

We have created a computational approach to detect drug resistance in IAV sequences by applying N-gram analysis and machine learning algorithms on data acquired from IRD. N-grams have been successfully used for a long time in a wide variety of problems and domains including language identification, text categorization, optical character recognition and music categorization. N-grams have also been more recently applied to problems in computational biology. The method of N-grams was initially applied to large scale clustering of DNA texts [38]. Tomovic et al. computed dissimilarity functions of n-grams in HIV-1, HIV-2 and SHIV whole genome sequences to evaluate how accurately they could predict the family to which a sequence belonged, reaching an accuracy of 99% [39]. Strong correlations between n-gram patterns and secondary structure type have been identified based on 3D structures in PDB [40]. Tobi et al suggested that recruitment of rare 3-grams may be an efficient mechanism for increasing specificity at functional sites [41]. In a more recent study, n-gram analysis was used to classify species and to determine to what degree was the identity of the detected n-grams a property of phosphosites [42]. In [43], Iqbal et al. used a combination of 1, 2 and 3 grams to create a feature vector for classification and feature selection of protein sequence data.

From an information perspective, each DNA sequence is a linear text over the four-letter alphabet {A, C, G, T}. There are 256 possible 4-grams (n=4) and 64 possible 3-grams (n=3) for a given DNA sequence. Similarly, there are 8000 possible 3-grams for a given protein sequence. Our approach uses normalized frequency counts of 3-grams as feature vectors in a supervised machine learning model over datasets downloaded from IRD. The choice of 3-grams in this work is based on our experiments that showed that 1-grams and 2-grams do not confer enough specificity while there is no major difference in the performance of 3-grams and 4-grams (potentially since some strong signals characteristic of 3-grams may be overlooked upon examination of 4-grams).

In this study, we performed binary classification of input proteins and nucleotide sequences using multiple machine learning algorithms. Sequences are classified using 4 supervised classification approaches: K-Nearest neighbors, Gaussian Naïve Bayes, Support vector machine (linear kernel) and Random Forest. We use the same normalized n-gram frequency counts as input for all our classification models. We could classify both nucleotide and protein sequences with a high degree of accuracy. In this paper, we have presented details of our experiments for our protein data sets and we include a snapshot of our results with nucleotide sequences.

Since the relative rank of a feature (n-gram) used as a decision node in a tree-based classification task can be used to assess the relative importance of that n-gram with respect to the classification, we performed "feature selection" based on the random forest classifier [44] to extract important n-grams to get insights into residue positions that are most responsible for classification accuracy. These residue positions are particularly interesting for our experiments with "drug resistance" since they provide us valuable information regarding potential mutations that are responsible for drug resistance.

When we observed a high degree of accuracy in classifying disease resistance of proteins that are not directly targeted by a specific drug (Adamantanes, Oseltamivir), we conducted correlation analysis using a relatively new correlation measure called Maximal Information Coefficient (MIC) [27] to identify correlations among mutations between proteins in IAV. MIC is particularly useful for identifying linear and non-linear relationships in data with high dimensionality and is finding increasing applicability in bioinformatics [45], [46]. We were specifically interested in identifying point mutations in M2 and NA with high degree of association with mutations in HA and other proteins.

Methods

Data

DNA and Protein sequences used in this work have been downloaded from Influenza Research Database (IRD). Non-redundant Sequences of IAV that are associated with specific characteristics (drug resistance, similarity to 2009 pandemic H1N1, geography, year, sub-type) have been downloaded after de-duplication. Details of specific data sets of protein sequences downloaded for classification are presented in Table 1.

Table 1 - Datasets used for Classification

No	Classification	Classes	Protein	Sequence Count
1	Viral subtypes	H1N1, H3N2	HA	500
2	Viral subtypes	H1N1, H3N2	NA	486
3	Viral subtypes	H1N1, H3N2	M1	365
4	Viral subtypes	H1N1, H3N2	NP	500
5	Drug resistance	Adamantane Resistant, Adamantane Sensitive	HA	228
6	Drug resistance	Adamantane Resistant, Adamantane Sensitive	M2	28
7	Drug resistance	Oseltamavir Resistant, Oseltamavir Sensitive	HA	189
8	Drug resistance	Oseltamavir Resistant, Oseltamavir Sensitive	NA	98
9	Drug resistance	Adamantane Resistant and Oseltamavir Sensitive, Adamantane Sensitive and Oseltamavir Resistant	HA	180
10	Drug resistance	Adamantane Resistant and Oseltamavir Sensitive, Adamantane Sensitive and Oseltamavir Resistant	NA	90
11	Drug resistance	Adamantane Resistant and Oseltamavir Sensitive, Adamantane Sensitive and Oseltamavir Resistant	M1	16

12	Drug resistance	Adamantane Resistant and Oseltamavir Sensitive, Adamantane Sensitive and Oseltamavir Resistant	M2	18
13	Similarity to 2009 pH1N1	pH1N1, Not-pH1N1	HA	500
14	Similarity to 2009 pH1N1	pH1N1, Not-pH1N1	NA	500
15	Similarity to 2009 pH1N1	pH1N1, Not-pH1N1	M1	500
16	Similarity to 2009 pH1N1	pH1N1, Not-pH1N1	PB1	500
17	Geography	Asia, North America	HA	500
18	Geography	Asia, North America	NA	452
19	Geography	Asia, North America	NP	114
20	Flu season	2013-14, 2014-15	HA	418
21	Flu season	2013-14, 2014-15	NA	317
22	Flu season	2013-14, 2014-15	PA	50

Pipeline

Our implementation pipeline for classifying IAV sequences using n-grams and machine learning algorithms consisted of multiple steps. Python has been used as the programming language for implementation of this computational pipeline. This pipeline begins with a data cleansing and preparation step where sequences are parsed and extracted from two fasta files using features in biopython [47]. After appropriate cleansing and de-duplication, normalized 3-gram frequency matrices are computed for both the input files. Note that a single input fasta file consists of one class of data (listed in Table 1) and this pipeline does not expect aligned sequences. These 3-gram frequency counts are then used as feature vectors by classifiers in the next step as part of a training step to create a model that is iteratively tuned and finally tested using 10-fold cross validation. Four different classifiers based on Random Forest (RF), Support Vector Machine (SVM), K-Nearest

Neighbors (KNN) and Gaussian Naïve Bayes (GNB) algorithms are used to classify the data and results are depicted and compared using roc plots. 67% of input data has been used for training and the remaining 33% has been used for testing the model.

N-grams and feature extraction

After downloading appropriate DNA & protein sequence data from IRD, a python program computed 64 (normalized) 3-gram frequency counts for each of the DNA sequences and 8000 (normalized) 3-gram frequency counts for proteins sequences. These frequency counts are used as the feature vector for automated classification.

Automated classification

Automated binary classification of feature vectors was performed using python libraries in scikit-learn [48], [49]. Binary classification of feature vectors was separately conducted using four different classification methods – RF, SVM (linear kernel), KNN and GNB. A training, test split of (0.67, 0.33) was used for all the methods. We have used normalized frequency counts of 3-grams for all classifications. ROC plots have been created against the original test labels and predicted labels. A plotting package in python called matplotlib [50] has been used for creating ROC plots. Efficiency of the models were also evaluated using 10-fold cross validation.

Correlations

Adamantine and oseltamivir resistance in IAV strains was mainly attributed to mutations in M2 and NA proteins respectively. Our classification accuracy results for drug resistant and drug sensitive strains yielded a high accuracy rate for proteins other than NA and M2 and this led us to explore potential correlations between mutations in NA/M2 and other proteins. We have used a python implementation of MIC called minepy [51] to calculate pair-wise correlations in mutations between residues in two proteins. The overall pipeline for computing MIC-based correlations is comprised of the following steps. First, we downloaded aligned protein sequences of HA, M1, M2, NA and NP for H1N1 strains that are "adamantine resistant and oseltamivir sensitive" in fasta format from IRD. Sequence combinations of two proteins from the above set of proteins ([HA, M1], [HA, M2], [HA, NA], [HA, NP], [M1, M2]) were read by a python program. Strains that contained sequences for one protein but not the other are discarded and only strains with sequences from both the proteins are considered for downstream processing. For each position i of a given protein (in the two proteins), the type of amino acid s of the multiple sequence alignment is represented by a binary variable xi(s) where xi(s) = 1 if the amino-acid is the most frequent amino acid at this position within the MSA, and xi(s) = 0 if it is another amino acid. Inter-residue pair-wise maximal information coefficient was calculated between residues in the two proteins using minepy. As an example, MIC scores for each of the 568 positions in HA are evaluated with respect to each of the 254 positions in M1.

Classification Results

To classify IAV sequences, we executed our pipeline for several different data sets. The experiments were arranged into multiple groups based on the specific classification type: year of infection, subtype, geography, similarity to 2009 pandemic H1N1 virus and drug resistance. In each of these experiments, protein and nucleotide sequences from IRD were used as data sets. Results for protein and nucleotide data sets are included in this section. We primarily focused on binary classification problems in our experiments. Based on results from our preliminary tests, we used 3-grams in all our experiments. Efficiency of our models was evaluated using 10-fold cross validation, and (balanced) accuracy results of our classification runs were graphically depicted.

Drug Resistance

Classification of strains based on known drug resistance attributes was performed with accuracy results nearing 100% with all four methods. We performed three separate experiments to perform classification based on drug resistance. In our first experiment, we classified adamantine resistant and adamantine sensitive strains of M2 and HA proteins. In our second experiment, we classified oseltamivir resistant and oseltamivir sensitive sequences of NA and HA and for our final experiment, we attempted to classify 'adamantane resistant and oseltamivir sensitive' and 'adamantane sensitive and oseltamivir resistant' sequences of HA, NA, M1 and M2 proteins. These experiments resulted in a near-100% classification accuracy.

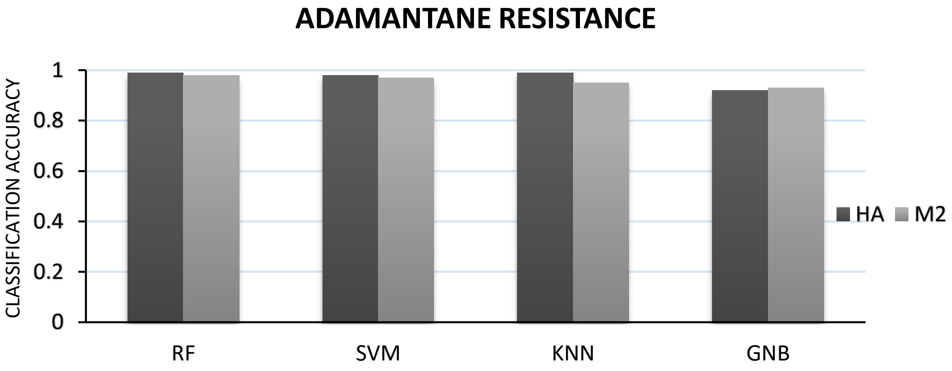


Figure 3 - Classification accuracy for adamantane resistant vs adamantane sensitive sequences of HA and M2 using four different classification algorithms

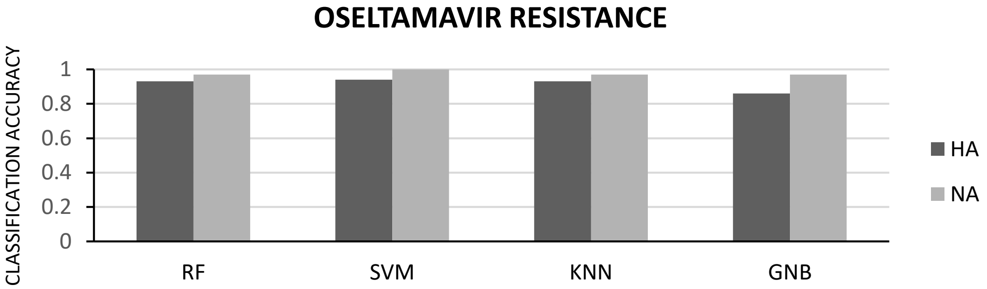


Figure 4 - Classification accuracy for oseltamavir resistant vs oseltamavir sensitive sequences of HA and NA using four different classification algorithms

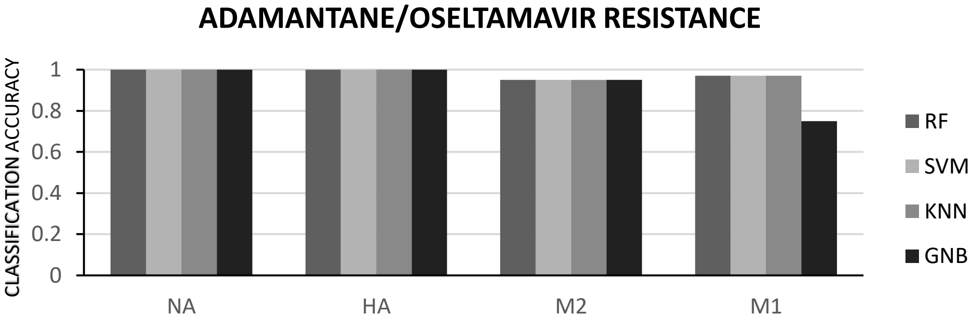


Figure 5 - Classification accuracy for adamantane-resistant, oseltamavir-sensitive vs adamantane-sensitive, oseltamavir-resistant sequences of HA, NA, M1& M2 using four different classification algorithms

To understand the most significant 3-grams that are responsible for this clear split, we used random forest classifier to perform feature selection. Since the anti-viral activity of Adamantane has long been associated with its ability to bind and block the ion channel protein M2, we were particularly interested in the features used for classifying the M2 protein sequences. Table 2 provides a list of these 3-grams along with the most common location of the 3-gram in M2 sequence and clearly highlights that the features that are used by ‘random forest’ classifier are closely related to known drug-resistant mutations in M2. Most notably, our top 10 features contained multiple 3-grams that contained residue #31 (which is the site for S31N mutation that is known to be the most significant for drug resistance). In a similar manner, we performed feature selection for Oseltamavir resistant and Oseltamavir sensitive strains of NA, since Oseltamavir is known to be a NA inhibitor. Again, our top 10 features (listed in Table 3) contained multiple 3-grams that contained residue #275 (which is the site for H275Y mutation that is linked to drug resistance). This link between significant 3-grams and known mutations in M2 and NA validates our overall approach.

Table 2 - Important features in M2 (based on AD resistance vs AD sensitivity)

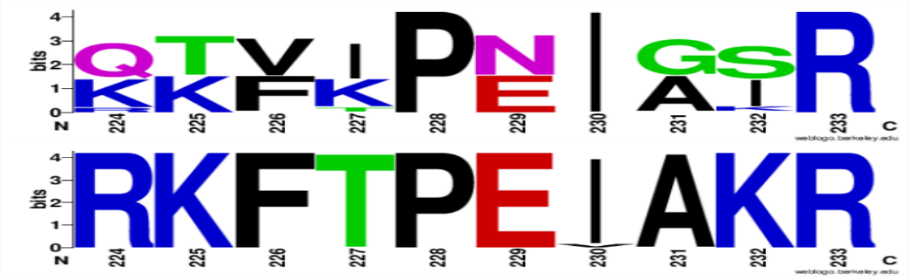
No.	3-gram	Position in M2 sequence	Known Mutations
1	ANI	30,31,32	S31N, A30T
2	YRE	76,77,78	
3	NII	31,32,33	S31N
4	SII	31,32,33	S31N
5	GIV	34,35,36	G34E
6	LHL	36,37,38	L38F
7	LFS	46,47,48	
8	GIL	34,35,36	G34E
9	FKC	48,49,50	
10	VHL	36,37,38	L38F

Table 3 - Important features in NA (based on OS resistance vs OS sensitivity)

No.	3-gram	Position in NA sequence	Known Mutations
1	YYE	275,276,277	H275Y
2	NFY	273,274,275	H275Y
3	SIE	266,267,268	
4	IVM	287,288,289	
5	GEG	331,332,33	
6	DGM	186,187,188	
7	MGW	188,189,190	
8	HYE	275,276,277	H275Y
9	FYY	274,275,276	H275Y
10	NQR	50,51,52	

To visualize these mutations, we created weblogos [52] of sequences around the position of interest. These logos depicted in Figure 6, Figure 7 & Figure 8 provide a good visual illustration of the differences between the drug sensitive and drug resistant proteins around the most significant 3-gram.

Figure 6 - weblogo for a 10-residue segment around the most significant 3-gram (227/228/229) in HA



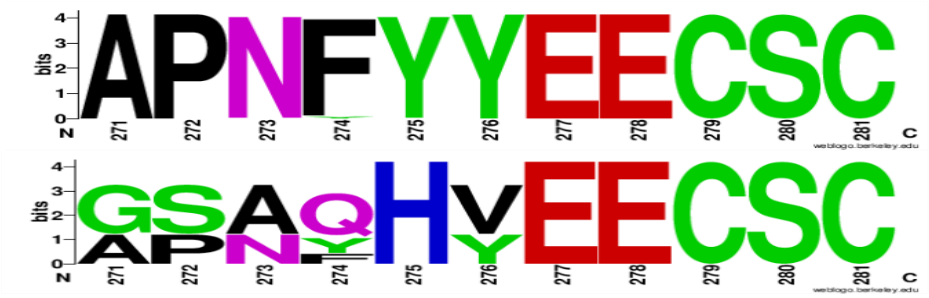


Figure 7 - weblogo for a 11-residue segment around the most significant 3-gram (274/275/276) in NA

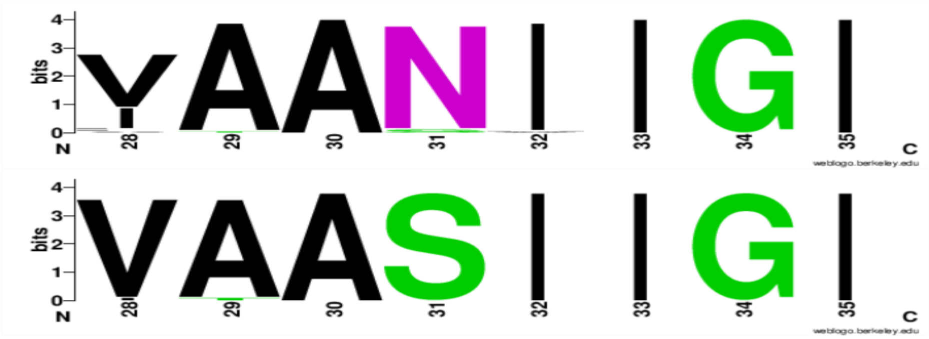


Figure 8 - weblogo for an 8-residue segment around the most significant 3-gram in M2

To further confirm the accuracy of our results, we created a control data set by shuffling the frequency counts for each sequence. ROC plots created using original datasets and control (random) datasets (depicted in Figure 9 and Figure 10) show an AUC close to 1 for original datasets and much lower AUC values for control datasets, indicating significance of our results. We performed a similar validation by shuffling the labels (instead of the feature vectors) of the two classes and observed a similar decrease in AUC.

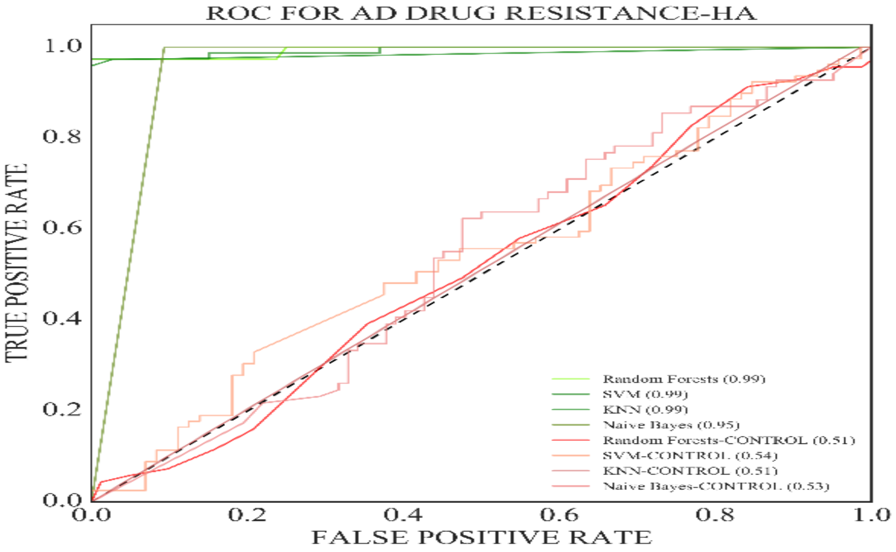


Figure 9 - ROC plots based on classification of AD resistant and AD sensitive sequences of HA, values in parenthesis represent area under curve (AUC)

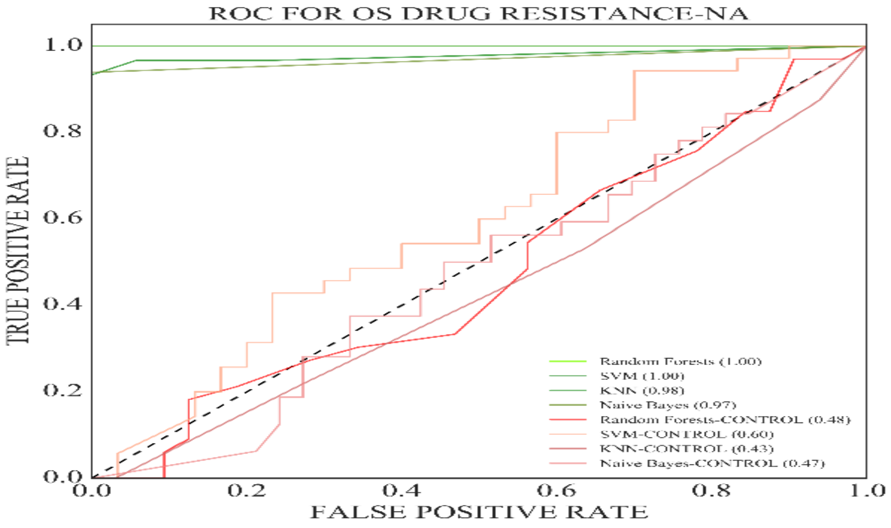


Figure 10 - ROC plots based on classification of OS resistant and OS sensitive sequences of NA, values in parenthesis represent area under curve (AUC)

Viral Subtypes

Classification of IAV sequences into viral subtypes H1N1 and H3N2, resulted in a clear split with classification rates consistently approaching 100% with all four classification algorithms for HA, NA, NP and M1 proteins. These accuracy results are depicted in Figure 11.

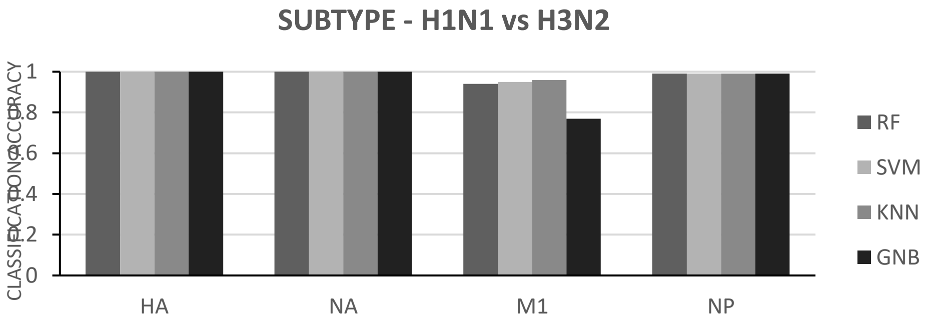


Figure 11 - Classification accuracy for H1N1 vs H3N2 classification

Geography

10-fold cross-validation accuracy for classification of IAV sequences of HA, NA and NP proteins based on geographic location (North America & Asia) ranged between 60% and 80% indicating that there is substantial difference between results based on the classification algorithm. To reduce noise, we selected sequences of H1N1 subtype and sequences from years 2011 to 2014 for both data sets. Classification accuracy was lower for NP compared to HA and NA proteins. These accuracy results are depicted in Figure 12. Based on these results, it can be inferred that there are important differences in sequences (of HA and NA) in circulation among humans in Asia and North America.

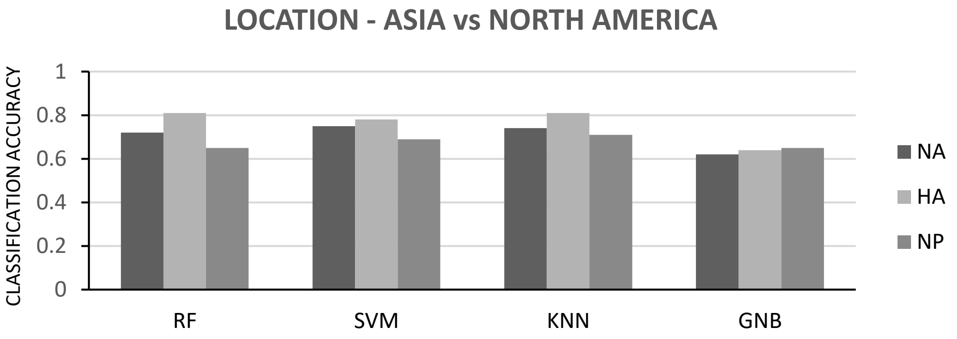


Figure 12 - Classification accuracy based on geographic location

Flu season

Classification between strains of IAV from a specific geographic location for two consecutive flu seasons (2013-2014 vs 2014-2015) proved to be moderately accurate confirming a known hypothesis (that there are substantial mutations in sequences of IAV from year to year). Results for HA and NA are better when compared to PA, again confirming our existing understanding that there are more mutations in surface proteins.

ROC curves generated by the classification yielded good results with an area close to 1 under the curve. These results correlate well with IAV's ability to swiftly mutate between consecutive flu seasons.

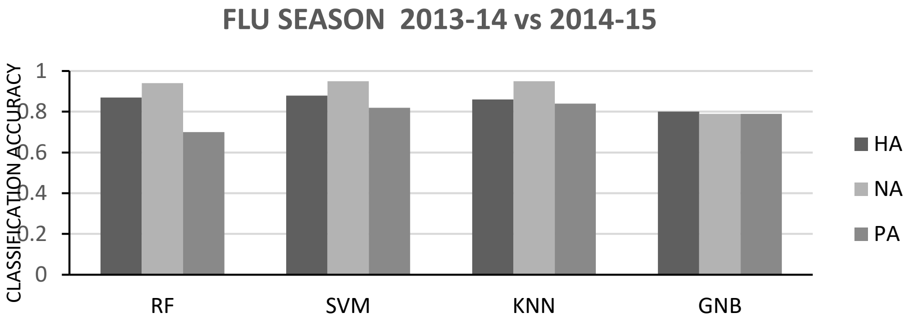


Figure 13 - Classification accuracy for 2013-14 vs 2014-15 sequences

Similarity to 2009 pH1N1

Classification of strains based on similarity to 2009 pandemic H1N1 was performed with an accuracy rate close to 100% for HA, NA, PB1 and M1 sequences. These results are depicted in Figure 14 for HA, NA, M1 and PB1 proteins.

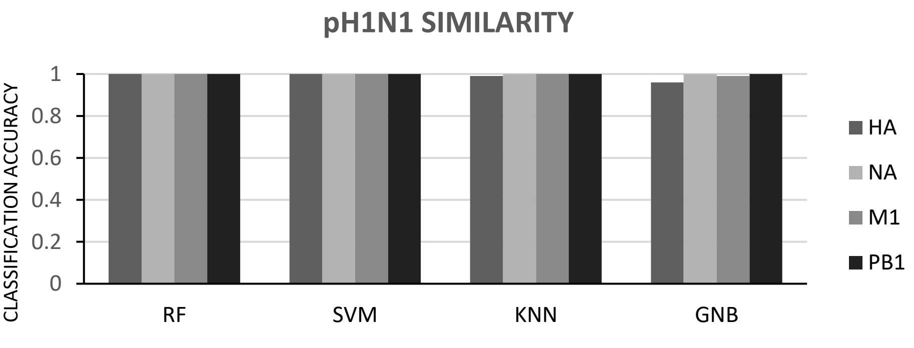


Figure 14 - Classification accuracy for protein sequences based on pH1N1 similarity

Classification of DNA sequences

We were also able to classify DNA sequences with a high degree of accuracy using the same pipeline elucidated in section 3.2. We fed (normalized) 3-gram counts as feature vectors to our classification models and could perform binary classification with similar accuracy levels as in protein sequence based classification. Classification accuracies for distinguishing sequences based on drug resistance and flu season are depicted in Figure 15 and Figure 16 respectively.

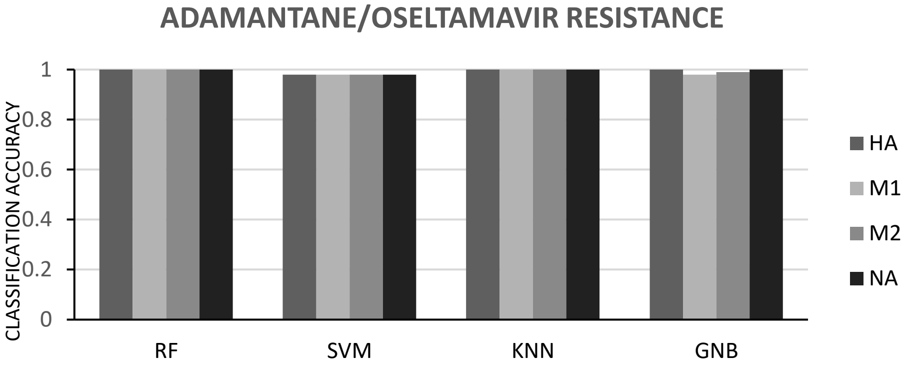


Figure 15 - Classification accuracy for adamantane-resistant, oseltamavir-sensitive vs adamantane-sensitive, oseltamavir-resistant DNA sequences of HA, NA, M1& M2 using four different classification algorithms

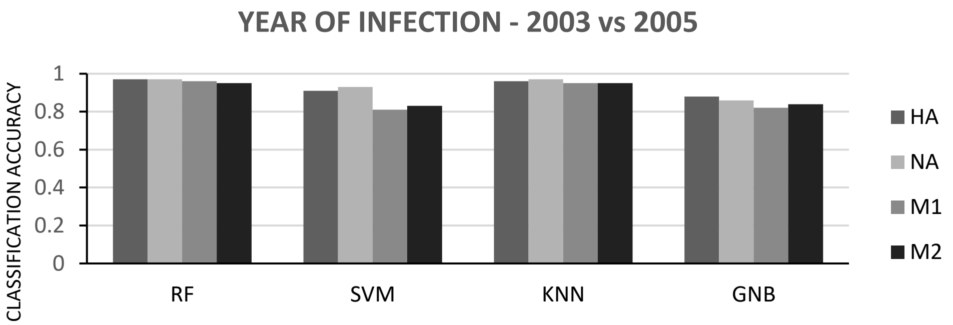


Figure 16 - Classification accuracy based on year of infection for HA, NA, M1& M2 DNA sequences using four different classification algorithms

Classification of protein sequences using a reduced 3-letter alphabet

We have also attempted to classify IAV protein sequences using reduced 3-letter alphabets detailed in [53], [54]. We modified our pipeline by adding an additional step to convert the protein sequences to reduced (3-letter) alphabets and computed a feature vector comprising 27 normalized 3-gram frequency counts. Our classification results were at or below 50% for both the original and control data for all the 22 data sets listed in Table 1 suggesting loss of important differentiating information when we converted the original protein sequences to 3-letter alphabet.

Correlations

Approximately 300 FASTA aligned sequences of adamantine resistant and oseltamavir sensitive strains of HA, M1, M2, NA and NP sequences are downloaded from IRD. Sequence lengths of HA, NP, NA, M1, M2 proteins are 568, 500, 472, 254 and 99 respectively. Inter-residue 'Maximal Information Coefficient' is calculated for residues in all possible combinations of two proteins from the above set of proteins ([HA, M1], [HA, M2], [HA, NA], [HA, NP], [M1, M2]).

Figure 17 and Figure 18 provide two visualizations of the MIC based correlation scores. In these plots, MIC scores are plotted using a sequential color map where there is a clear progression from lighter shades of grey (for lower MIC values) to brighter shades of grey (for higher MIC values). Figure 17 depicts pair-wise correlation scores for (NA, HA) and (NA, M1) proteins and Figure 18 provides a similar visualization for (HA, M1) and (HA, M2) proteins. These plots clearly illustrate that there are several correlated mutations between these proteins and confirm our hypothesis that these correlations among mutations can potentially be the reason for our ability to classify drug resistance in proteins that are not directly targeted by the drug.

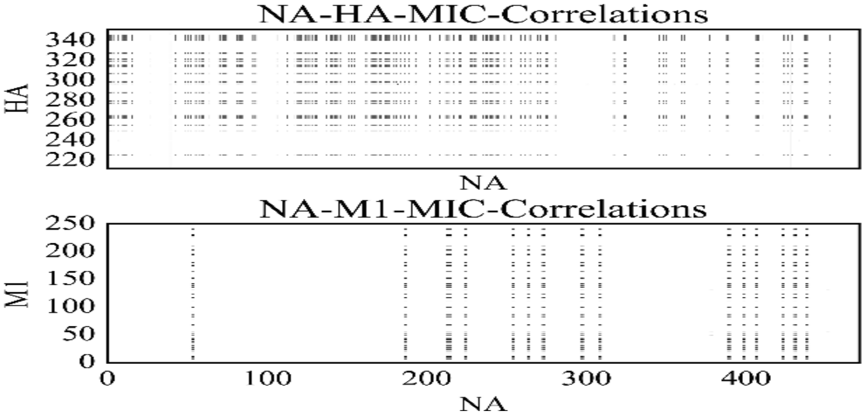


Figure 17 - MIC inter-residue correlation scores between NA, HA and NA, M1 proteins. x and y axis represent residue numbers within that protein

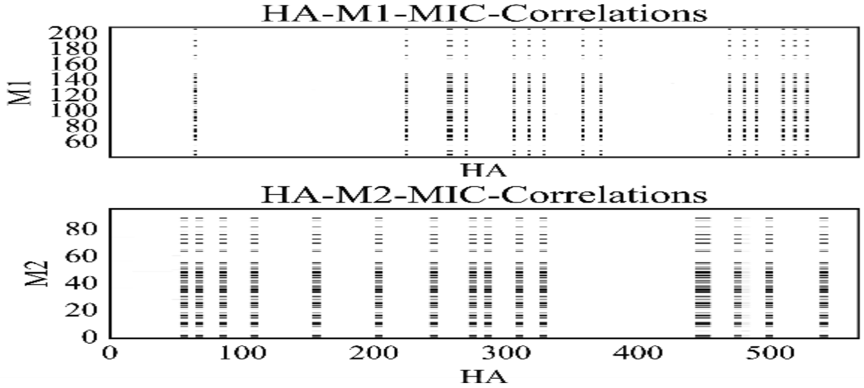


Figure 18 - MIC inter-residue correlation scores between NA, H1 and NA, M1 proteins. x and y axis represent residue numbers within that protein

Discussion

N-gram analysis coupled with supervised classification algorithms to distinguish between strains of IAV proved to be successful. We created a software pipeline in Python and applied it to classify protein and nucleotide sequences of IAV that we downloaded from IRD. Using this approach, we could perform binary classification to distinguish sequences from different subtypes, (consecutive) flu seasons, geographic locations (Asia vs North-America), similarity to 2009 pandemic H1N1 and most importantly drug resistance. We could classify sequences of HA, NA, M1, M2 and NA based on their resistance to Adamantane and Oseltamavir with a near 100% accuracy. Using Random Forest classifier, we identified the most significant features (3-grams) in NA and M2 sequences and could confirm a strong linkage between these features and known drug resistant mutations in these proteins.

CHAPTER 4 - LITERATURE REVIEW

In this chapter, we will cover details of prior related research. Our literature review can be broadly classified into three main areas –

1. Analysis of correlated mutations in genomic and proteomic datasets
2. Application of network analysis to understand structural dynamics of biological systems
3. Other relevant literature covering topics like
    a. Application of Maximal Information Coefficient (MIC) to understand non-linear associations in biological datasets
    b. Use of "correlation coefficient" as a weight measure in creation of networks
    c. Vaccines, Anti-viral drugs and Epidemiology of Influenza

Correlated Mutations

Several studies have been conducted up to date, to understand correlated mutations within proteins. The tendency for correlated mutations to indicate contacts between two or more residues was originally reported in 1994 [5]. In [55], authors show that a direct contact is more likely to be present when the correlation between the positions is strong at the amino acid level but weak at the codon level. Most methods that predict protein contacts between residues based on correlated mutations tend to have a high false positive rate. In [56], authors reported a new implementation where selection rules are applied to improve the overall accuracy. In 2013, Tayler et al. [57] conducted a thorough review of research conducted in the field of contact prediction from correlated substitutions and also discussed applications of prior work to protein and RNA structure prediction. In [58], authors report a novel method called PSICOV that uses sparse inverse covariance estimation to perform protein contact prediction. In [59], Weigt et al. employed a combination of covariance and global inference approaches to successfully identify directly correlated residue pairs.

Apart from contact prediction, there were several efforts that have been reported that focused primarily on algorithms and methods to identify correlated mutations. One of the earlier works [60] employed a conditional approach using 2*2 frequency tables, to detect correlated mutations in V3 loop of the envelope gene from human immunodeficiency virus (HIV). Fares et al. created a tool called CAPS (coevolution analysis in protein sequences) that identifies co-evolution between amino acid sites in a protein sequence using blosum distance measures [61]. In [62], Wang and Lee elucidated a computational methodology based on analysis of synonymous and non-synonymous mutations to distinguish background linkage disequilibrium (LD) from covariation due to selection pressure. In 2008, Correlated mutations in HIV-1 protease were analyzed using spectral clustering of covariance matrices [63]. In a 2009 study [64], Andrec et al. developed a probabilistic approach based on degree of connected information, to identify second and higher order correlations within HIV-1 protease. Mao et al. provided a comprehensive comparison of existing approaches for detecting coevolution [65].

Majority of studies done up to date in this area have focused on residue contact prediction and structural implications of correlated mutations and not much work has been reported on potential phenotypic implications. Kowarch et al [8] have conducted an analysis of disease causing correlated mutations in humans to prove their hypothesis that conservation and co-evolution of residues in a protein influences the likelihood of a residue to be functionally important.

In [66], Hu analyzed sequences from 2009 H1N1 Influenza pandemic, and identified two networks of co-mutations that may potentially affect the flu-drug binding sites on neuraminidase (NA). Hu also explored host differentiation and co-mutations in M, NS and PB1 of avian, human, 2009 H1N1 and swine viruses with random forests, information entropy and mutual information [67]. In [68] Hu analyzed co-mutated sites within and between four important proteins – NP, PA, PB1 and PB2 of avian, human, pandemic 2009 H1N1, and swine flu using mutual information, based on which several highly connected networks of correlated sites in NP, PA, PB1 and PB2 were discovered. [69] employed concepts from 'random matrix theory' to determine collectively coevolving groups of residues in HIV Gag polyproteins. It has also been demonstrated in a more recent work that evolution in human Influenza A is mainly driven by dynamically correlated mutations [70]. [71] performed evolution analyses to illuminate dependencies between amino acid sites in the chaperonin system GroES-L. In [72], Giuliani et al. demonstrated the possibility to predict correlated mutations in a single protein system (Hepatitis C Virus NS5B viral RNA polymerase protein) using a combination of supervised (discriminant analysis) and unsupervised (principal component analysis) approaches.

During our literature survey, we have seen very few efforts where computational tools and/or web applications have been developed to help detection and analysis of covarying substitutions. Interprotein Correlated Mutation Server (ICOMS) was developed to estimate covariation between residues of different proteins by using 4 different covariation methods [73]. In a separate effort, Li et al. developed a stand-alone R/Bioconductor package called CorMut to detect the correlated mutations among positive selection sites [74].

Network Analysis

The concepts of network theory have been applied to the study of protein structure and function with promising results.

Protein-protein interaction networks, regulatory networks and signal transduction networks (collectively termed as Protein networks) are typically designed to model interactions between proteins or other macro molecules while Protein structure networks represent interactions between segments of a protein [75]. Majority of studies conducted up to date in protein structure networks focused on analysis of amino acid interactions where the network elements are amino acids and edges are created between two nodes if the distance between them is below a given threshold.

We have reviewed several works where network analysis was applied to model biological networks and more specifically literature where concepts of correlation have been applied to model interactions between biological elements. In [76], authors reported an integrative strategy combining quantitative genetic mapping and metabolite-transcript correlation networks to identify functional associations in Arabidopsis Thaliana. Batushansky et al. [77] introduced a series of methods for correlation based network generation and analysis using freely available software, and applied their methods on metabolomics data of a population of human breast carcinoma cell lines. In a more recent study conducted by Toubiana et al., [78], Correlation-based Network Analysis (CNA) was conducted to investigate the natural variability of leaf metabolism and enzymatic activity in a maize inbred population. In most of these biological networks created from CNA, vertices typically represented molecular elements (example: genes, proteins, metabolites) and edges represented a correlation coefficient between the elements.

The other category of work that we reviewed was in the area of protein structure networks where residues in proteins act as vertices and we noticed that most of the efforts relied on energy or distance measures to create edges. [79] studied protein stability by constructing a network of non-covalent connections between amino acid side chains. In [75], Bode et al. reviewed results of topological analysis of protein structures as molecular networks describing their small-world non scale-free character. Estrada [80] proposed a modification of Watts-Strogatz model [81] to describe protein residue networks. He studied 595 non-homologous proteins and concluded that they exhibit universal topological characteristics. In their model, Gaci and Balev [82] created edges between amino acids if the distance between C_α atoms is less than 7Å and they identified a number of general properties of these distance-based networks. They also introduced the term "Amino-acid Interaction Networks" to broadly describe networks that are created based on interactions between amino acids. In a subsequent research publication [83], Gaci provided a topological description of hubs in Amino-acid Interaction Networks. *We have seen very few previous studies where network analysis was conducted based on correlations between residues, and more specifically we did not come across any efforts that focused on comprehensive network analysis based on correlated mutations among residues in proteins* *in Influenza*. A study published in 2008 by Du et al. [84] attempted to understand the dynamics of H3N2 evolution by constructing and analyzing nucleotide co-occurrence networks. Hu’s analysis reported in 2009 [68] used ‘Mutual Information’ to quantify correlation of sites within 4 Influenza proteins (NP, PA, PB1 and PB2) and uncovered interaction patterns but did not apply broader principles from graph theory and network analysis.

CHAPTER 5 - DATASETS

We have downloaded several datasets from Influenza Research Database (IRD) (http://www.fludb.org) [29] during the course of this dissertation. The IRD is a free, open, publicly-accessible resource funded by the U.S. National Institute of Allergy and Infectious Diseases through the Bioinformatics Resource Centers program. IRD provides a comprehensive, integrated database and analysis resources for Influenza sequence, surveillance, and research data, including user-friendly interfaces for data retrieval, visualization, and comparative genomics analysis, together with personal log in-protected 'workbench' spaces for saving data sets and analysis results. In this chapter, we provide an overview of options we picked and steps we followed to download the data along with few screenshots. We also provide a listing of different datasets.

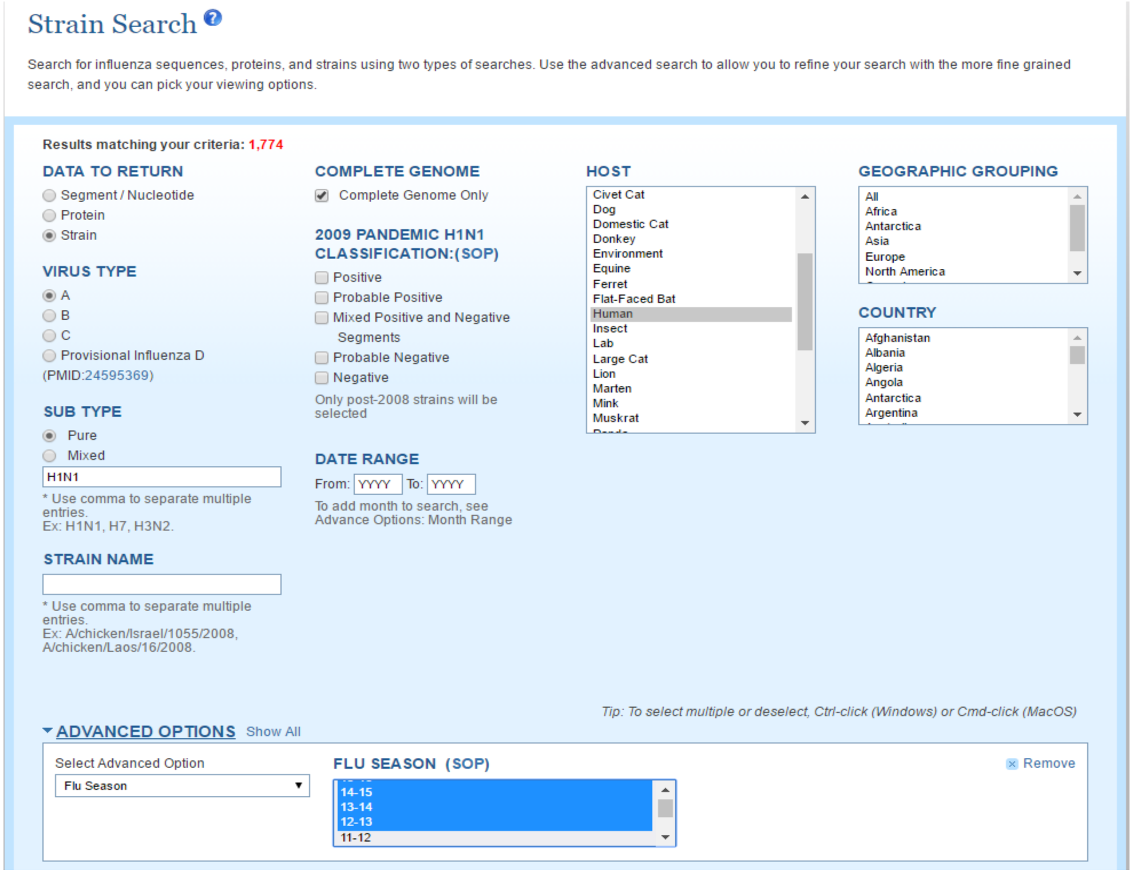
Workflow

We have only used "strain" data from IRD to perform network analysis of correlated mutations. For each dataset that we downloaded, this is the sequence of steps we followed.

1. We searched for 'strains' using the *Search Data* → *Search Sequences* → *Strain Data* from IRD's main menu.
2. Based on the specific dataset that we are interested, we selected specific options that are generally depicted in Figure 19. For all our datasets, we have always selected strains of Influenza A with complete genome only and we limited our search results to strains from a pure sub-type (excluding all

mixed sub-types). We have not limited our search results to either a geographic grouping or a country. The ‘sub type’, ‘host’ and ‘flu season’ varied for each of our datasets.

3. After selecting appropriate options, we submit our search and select the results in the following screen. We download the strains in ‘Protein FASTA’ format. This will result in a single FASTA file that contains sequences of all the 10 proteins that we are interested in. Figure 20 and Figure 21 depict these steps of the workflow.



Strain Search

Search for influenza sequences, proteins, and strains using two types of searches. Use the advanced search to allow you to refine your search with the more fine grained search, and you can pick your viewing options.

Results matching your criteria: 1,774

DATA TO RETURN
○ Segment / Nucleotide
○ Protein
● Strain

VIRUS TYPE
● A
○ B
○ C
○ Provisional Influenza D (PMID:24595369)

SUB TYPE
● Pure
○ Mixed
H1N1
* Use comma to separate multiple entries.
Ex: H1N1, H7, H3N2.

STRAIN NAME

* Use comma to separate multiple entries.
Ex: A/chicken/Israel/1055/2008, A/chicken/Laos/16/2008.

COMPLETE GENOME
☑ Complete Genome Only

2009 PANDEMIC H1N1 CLASSIFICATION:(SOP)
□ Positive
□ Probable Positive
□ Mixed Positive and Negative Segments
□ Probable Negative
□ Negative
Only post-2008 strains will be selected

DATE RANGE
From: YYYY To: YYYY
To add month to search, see Advance Options: Month Range

HOST
Civet Cat
Dog
Domestic Cat
Donkey
Environment
Equine
Ferret
Flat-Faced Bat
Human
Insect
Lab
Large Cat
Lion
Marten
Mink
Muskrat

GEOGRAPHIC GROUPING
All
Africa
Antarctica
Asia
Europe
North America

COUNTRY
Afghanistan
Albania
Algeria
Angola
Antarctica
Argentina

Tip: To select multiple or deselect, Ctrl-click (Windows) or Cmd-click (MacOS)

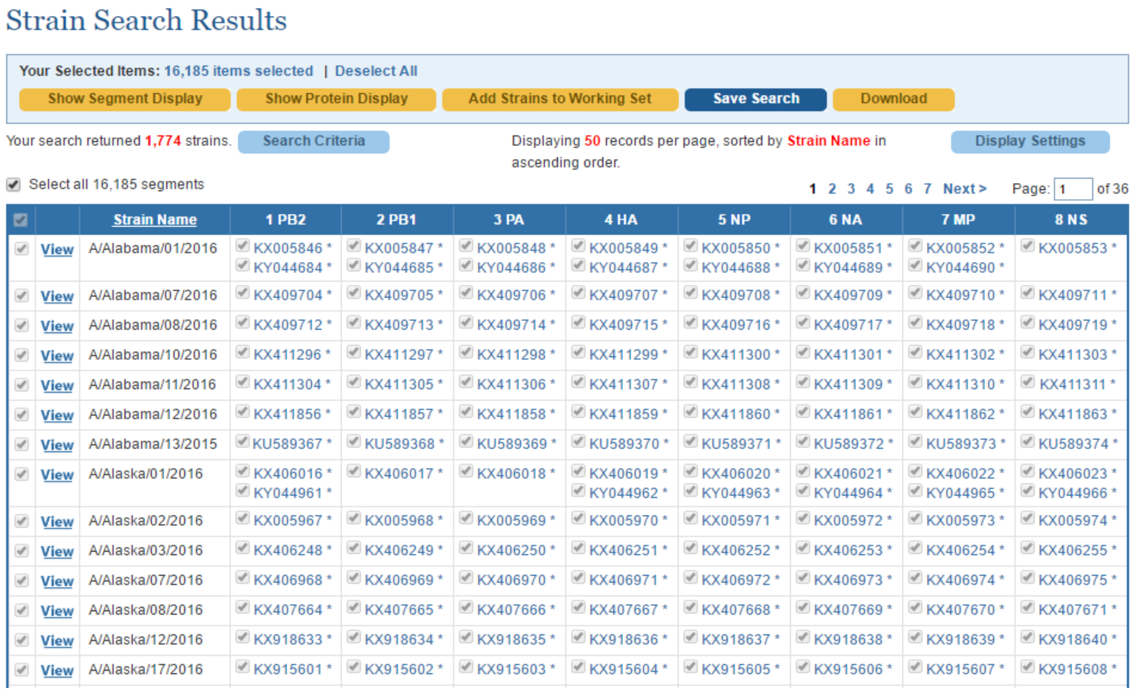
ADVANCED OPTIONS Show All

Select Advanced Option
Flu Season

FLU SEASON (SOP)
14-15
13-14
12-13
11-12

Remove

Figure 19 - Search options in the IRD



Strain Search Results

Your Selected Items: 16,185 items selected | Deselect All

Your search returned 1,774 strains.

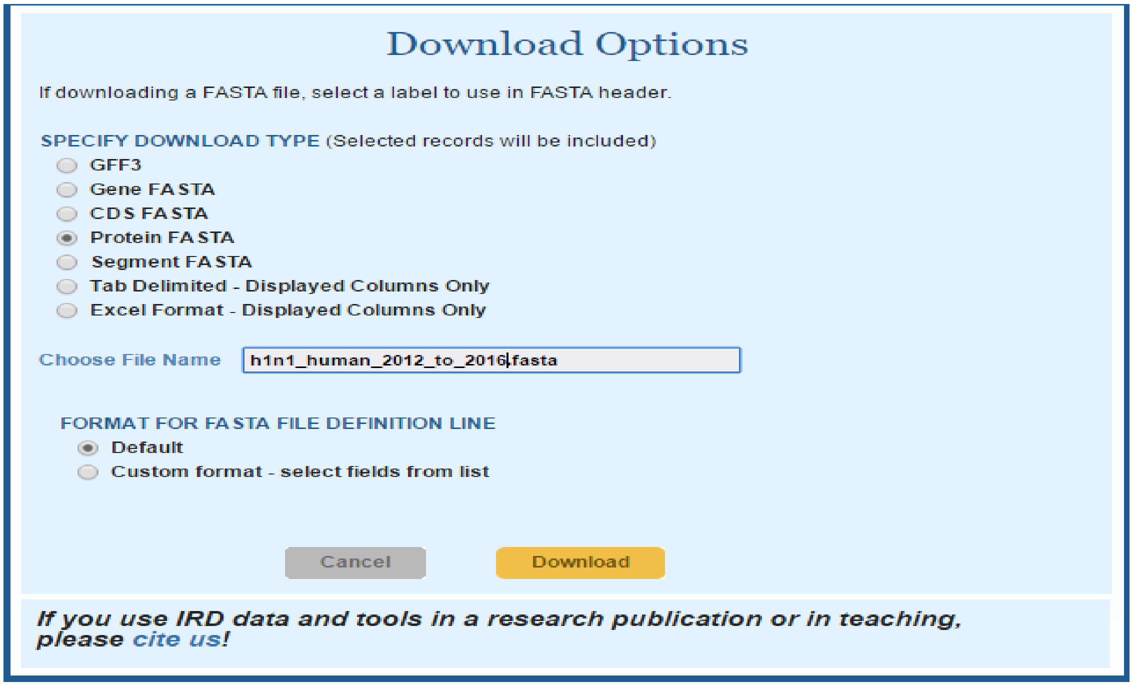
Displaying 50 records per page, sorted by Strain Name in ascending order.

☑ Select all 16,185 segments

Page: 1 of 36

☑		Strain Name	1 PB2	2 PB1	3 PA	4 HA	5 NP	6 NA	7 MP	8 NS
☑	View	A/Alabama/01/2016	☑ KX005846 * ☑ KY044684 *	☑ KX005847 * ☑ KY044685 *	☑ KX005848 * ☑ KY044686 *	☑ KX005849 * ☑ KY044687 *	☑ KX005850 * ☑ KY044688 *	☑ KX005851 * ☑ KY044689 *	☑ KX005852 * ☑ KY044690 *	☑ KX005853 *
☑	View	A/Alabama/07/2016	☑ KX409704 *	☑ KX409705 *	☑ KX409706 *	☑ KX409707 *	☑ KX409708 *	☑ KX409709 *	☑ KX409710 *	☑ KX409711 *
☑	View	A/Alabama/08/2016	☑ KX409712 *	☑ KX409713 *	☑ KX409714 *	☑ KX409715 *	☑ KX409716 *	☑ KX409717 *	☑ KX409718 *	☑ KX409719 *
☑	View	A/Alabama/10/2016	☑ KX411296 *	☑ KX411297 *	☑ KX411298 *	☑ KX411299 *	☑ KX411300 *	☑ KX411301 *	☑ KX411302 *	☑ KX411303 *
☑	View	A/Alabama/11/2016	☑ KX411304 *	☑ KX411305 *	☑ KX411306 *	☑ KX411307 *	☑ KX411308 *	☑ KX411309 *	☑ KX411310 *	☑ KX411311 *
☑	View	A/Alabama/12/2016	☑ KX411856 *	☑ KX411857 *	☑ KX411858 *	☑ KX411859 *	☑ KX411860 *	☑ KX411861 *	☑ KX411862 *	☑ KX411863 *
☑	View	A/Alabama/13/2015	☑ KU589367 *	☑ KU589368 *	☑ KU589369 *	☑ KU589370 *	☑ KU589371 *	☑ KU589372 *	☑ KU589373 *	☑ KU589374 *
☑	View	A/Alaska/01/2016	☑ KX406016 * ☑ KY044961 *	☑ KX406017 *	☑ KX406018 *	☑ KX406019 * ☑ KY044962 *	☑ KX406020 * ☑ KY044963 *	☑ KX406021 * ☑ KY044964 *	☑ KX406022 * ☑ KY044965 *	☑ KX406023 * ☑ KY044966 *
☑	View	A/Alaska/02/2016	☑ KX005967 *	☑ KX005968 *	☑ KX005969 *	☑ KX005970 *	☑ KX005971 *	☑ KX005972 *	☑ KX005973 *	☑ KX005974 *
☑	View	A/Alaska/03/2016	☑ KX406248 *	☑ KX406249 *	☑ KX406250 *	☑ KX406251 *	☑ KX406252 *	☑ KX406253 *	☑ KX406254 *	☑ KX406255 *
☑	View	A/Alaska/07/2016	☑ KX406968 *	☑ KX406969 *	☑ KX406970 *	☑ KX406971 *	☑ KX406972 *	☑ KX406973 *	☑ KX406974 *	☑ KX406975 *
☑	View	A/Alaska/08/2016	☑ KX407664 *	☑ KX407665 *	☑ KX407666 *	☑ KX407667 *	☑ KX407668 *	☑ KX407669 *	☑ KX407670 *	☑ KX407671 *
☑	View	A/Alaska/12/2016	☑ KX918633 *	☑ KX918634 *	☑ KX918635 *	☑ KX918636 *	☑ KX918637 *	☑ KX918638 *	☑ KX918639 *	☑ KX918640 *
☑	View	A/Alaska/17/2016	☑ KX915601 *	☑ KX915602 *	☑ KX915603 *	☑ KX915604 *	☑ KX915605 *	☑ KX915606 *	☑ KX915607 *	☑ KX915608 *

Figure 20 - Search Results in the IRD



Download Options

If downloading a FASTA file, select a label to use in FASTA header.

SPECIFY DOWNLOAD TYPE (Selected records will be included)

○ GFF3
○ Gene FASTA
○ CDS FASTA
● Protein FASTA
○ Segment FASTA
○ Tab Delimited - Displayed Columns Only
○ Excel Format - Displayed Columns Only

Choose File Name h1n1_human_2012_to_2016.fasta

FORMAT FOR FASTA FILE DEFINITION LINE

● Default
○ Custom format - select fields from list

If you use IRD data and tools in a research publication or in teaching, please cite us!

Figure 21 - Download Results in the IRD

Details of Datasets

The primary datasets that we used in this work (listed in Table 4) were downloaded from the IRD using the workflow explained in previous section. The number of strains listed in this table is the number of unique strains after de-duplicating and the procedure we used to remove duplicate strains is explained in next chapter under *Preprocessing* section. Unless mentioned explicitly, these datasets have been downloaded in 01/2017.

Table 4 – Primary Datasets used for Network Analysis of Correlated Mutations

NAME	#STRAINS	COMMENTS
HUMAN_H1N1_ALL	1769	Human H1N1 strains from all years, maximum 300 strains per year
HUMAN_H3N2_ALL	1940	Human H3N2 strains from all years, maximum 300 strains per year

SWINE_H3N2_ALL	794	All available Swine H3N2 strains
SWINE_H1N1_ALL	1096	All available Swine H1N1 strains
AVIAN_H5_ALL	1463	All available Avian H5 strains
H7N9_ALL	434	All available H7N9 strains
HUMAN_ALL	9716	All available Human H1N1 strains downloaded in 10/2016

CHAPTER 6 - COMPUTATIONAL TOOLS AND LIBRARIES

Before we delve into the details of computational methodology and pipeline, we provide a listing of software tools and libraries used in our dissertation in this chapter to provide additional technical context. From a computational perspective, we have embraced the following tenets in choosing appropriate tools and technologies during this work.

1. We have used Python as the core programming language throughout the course of this work since Python provided the most robust and extensive set of open source libraries and tools for our end-to-end computational requirements that included data processing, scientific computations, network analysis, visualizations and web application. In a nutshell, the use of Python obviated the need to use multiple frameworks and/or programming languages during this work. At the same time, we have made meaningful exceptions to the use of Python wherever we felt that there is a gap in capabilities in the python eco system.
2. We have used only open-source tools and libraries during this work. This decision was made because there is sufficient quality and choice in the available tools and libraries in the open-source landscape, and to reduce the overall cost of the project.
3. We have attempted to reuse existing libraries wherever possible as opposed to building our own. From the very inception, we have acknowledged that this is an initiative that will focus on application of existing computational

methodologies and algorithms to biological datasets as opposed to developing such methodologies and/or algorithms. More importantly, our initial research quickly proved to us that there is sufficient maturity in the existing open-source software landscape that either meets or exceeds our requirements.

We created our data analysis pipeline using the following open-source tools and libraries based on the above guiding principles.

Table 5 - Computational Tools and Libraries

NAME	ADDITIONAL COMMENTS
Anaconda [85]	Anaconda is a distribution of Python for large-scale data processing, predictive analytics and scientific computing. It includes a collection of about 200 open source packages. Additional packages are available through contributed channels or through installation using a package manager. Anaconda obviates all issues surrounding installation of packages and use of multiple python versions and/or environments. For our work, we have used Anaconda 4.1.1 with Python 3.5.2.
Numpy [86]	Numeric Python (Numpy) is an open-source add-on module to Python. Numpy module provides pre-compiled basic mathematical and numerical routines for manipulating arrays and matrices of numeric data. We have used numpy version 1.11.1 in our work.
MUSCLE [87]	MUSCLE is a program for creating multiple alignments of amino acid or nucleotide sequences. A range of options is provided that give you the choice of optimizing accuracy, speed, or some compromise between the two. We have used 3.8.31 command-line version of MUSCLE to align sequences.
Neo4j [88]	Neo4j is a native ACID-compliant transactional graph database, designed by Neo Technology to store and process graphs from bottom to top. Neo4j graph database follows the ‘Property Graph Data Model’ to store and manage its data. Cypher is a declarative

	query language (originally created for Neo4j) that allows for expressive and efficient querying and updating of a property graph. In our dissertation, we have used neo4j community edition 3.1.0 to store IAV correlated mutation graphs and Cypher query language to query the database.
PyCharm [89], [90]	PyCharm is an Integrated Development Environment (IDE) for developing python applications. PyCharm's collection of out-of-the-box tools include an integrated debugger, code completion utility, test runner, integration with Git and PyCharm provides native support for python web development frameworks and JavaScript libraries. As part of this effort, we have used 2016.2.2 version of PyCharm (Professional Edition for Students)
Flask [91], [91], [92]	Flask is a micro framework for Python web applications. Flask is built with a small core and is easy to extend. We have used Flask (version 0.11.1) to create a web application to search and analyze correlated mutations in IAV strains.
Networkx [93], [94]	NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. While most of our graph related work was performed using Neo4j and Cypher, we have also used capabilities in Networkx (version 1.11) to create a graphml file for (import and) stand-alone analysis using Gephi.
Gephi [95]	Gephi is a stand-alone open-source network analysis and visualization tool. Gephi can be installed on both windows and linux platforms. After a graph is created (using an alternate library like networkx), Gephi allows us to import, analyze and visualize that graph.Gephi uses a 3D rendering engine to display large networks in real-time and to speed up the exploration. After creating graphml files of correlated mutation networks using networkx, we used Gephi (version 0.9.1) primarily to create compelling visualizations of the network and in some cases to perform preliminary analysis to understand network topology and view centrality distributions.
Bokeh [96]	Bokeh is a Python interactive visualization library that targets modern web browsers for presentation.

	We have used Bokeh (version 0.12.0) charts to create visualizations of node counts and edge counts vs MIC threshold values, for different datasets.
Matplotlib [97]	Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. We have used Matplotlib (version 1.5.1) plotting features to create macro plots to depict node and edge density for each of the 10 proteins.
py2neo [98]	Py2neo is a client library and toolkit for working with Neo4j from within Python applications. We have used Py2neo for both loading the graph database with node and relationship data, and for querying the database.
minepy [99]	Minepy is an open-source implementation of the MIC algorithm in python. As part of this dissertation, we performed all our MIC calculations using minepy to detect non-linear relationships between residues in proteins.
Biopython [47]	Biopython provides a collection of modules and scripts for developers of Python-based software for bioinformatics use and research. Biopython comes with reusable modules for parsing various bioinformatics file formats (BLAST, FASTA, Clustalw, PDB) and for interfacing with common programs used by the bioinformatics community. In this effort, we availed Biopython’s features for reading, parsing and writing alignments.
sigma.js [100]	Sigma is a JavaScript library dedicated to graph drawing. It makes easy to publish networks on web pages and allows developers to integrate network exploration in web applications. We created sigma.js visualizations of the entire network for each of our datasets to enable easy exploration of the graph.
Twitter Bootstrap	Bootstrap is a Cascading Stylesheets (CSS) framework that helps creation of sleek, intuitive front-ends for web applications.
scikit-learn [48]	scikit-learn provides a unified API for most supervised and unsupervised machine learning algorithms, along with utilities and meta-algorithms to help create an end-to-end machine learning pipeline to tie everything together.

CHAPTER 7 - IMPLEMENTATION DETAILS

In this chapter, we will provide details of the computational pipeline that we created to perform network analysis of correlated mutations in IAV strains. We will start with a high-level description of our methodology and central themes of our design. We will then provide a schematic of the main components before going into details of the pipeline for each of the components.

Core Themes

Before we go into specifics of our implementation, it is important to understand the core underlying themes of our computational methodology.

Modularity

First and foremost, we have focused on modularity of our overall solution and ensured that there are discrete reusable components in our pipeline with separate functions. This allowed us to test individual components before wiring them together to create an integrated solution. This approach also allowed us to try different techniques for implementation of specific components with ease.

Best of breed

Next important tenet that was integral to our methodology was to use a 'best of breed' approach when it comes to the choice of appropriate tools and technologies. From the choice of appropriate visualization libraries and/or tools to using the right paradigm for 'network analysis' and picking a simple (but sufficiently robust) web application framework, we adhered to a ‘best of breed’ philosophy taking our requirements into consideration.

Visualizations

Creation of compelling visualizations is key to any analysis. We have placed emphasis on depicting the results of our analysis using appropriate visualizations wherever appropriate.

Extensibility

From the very outset, we have ensured that the solution that we are designing will be extensible in nature and sufficiently generic. The ability to run the end-to-end pipeline for additional datasets without having to make changes to parts of the overall pipeline provided a degree of robustness to our solution and increased its applicability. While we started the initiative with datasets that are specific to Human IAV strains, we could quickly extend the solution to Swine and Avian datasets without additional complexity.

Graph Database

During this dissertation, we realized that there is tremendous value in storing our graphs in a database. This not only allowed us to separate the ‘graph creation’ aspect of our pipeline from subsequent analysis steps but also provided a flexible mechanism to extend the solution to additional datasets and data. We looked at available options and decided to use Neo4j graph database to store correlated mutation graphs.

Sequence Encoding

For us to compute correlations among mutations using a correlation coefficient like MIC, encoding the protein sequences to a numeric format is a required step. We have looked at several options for encoding and decided to go with a simplistic Boolean notation that was used in previous works [69]. While there are some inherent drawbacks in this approach (like the lack of differentiation between mutations that are more significant and less significant), this method tends to not be biased and is generic enough for our broad set of datasets where the specific context of a mutation is not known. Additional details of this notation are included in the section titled *MIC Computation*.

Web Application

We created a web application called NACMI (Network Analysis of Correlated Mutations in Influenza) to show results of our work. Functionality in this application can be broadly sub-divided into two categories. First, this application allows end-users to view topN nodes, edges and triplets in networks corresponding to individual datasets based on data available in a graph database, along with a feature to search for edges corresponding to specific residues in a protein. Second, this application provides static pre-computed visualizations of degree & clustering coefficient distributions, MIC distribution, entropy plots, interactive visualization of networks at 0.5 threshold and protein-level correlated networks. The combination of these two functionalities should allow users to comprehensively understand the overall characteristics of correlated mutation networks of IAV datasets.

Pipeline

Figure 22 provides a high-level schematic of our overall pipeline. We begin with a preprocessing step where we download and preprocess the data for downstream processing. The next step in our pipeline computes MIC correlations between residues in various proteins and generates CSV files with MIC scores. We then invoke a ‘graph creation’ sub-flow to create a correlated mutation graph. The ‘graph analysis’ sub-flow is responsible for creating necessary visualizations and performing appropriate analysis. As a final step, we perform post-processing to gain additional insights and to contextualize out results. We have designed and implemented our pipeline in a manner that allows us to run it consistently for all datasets that we download from IRD.

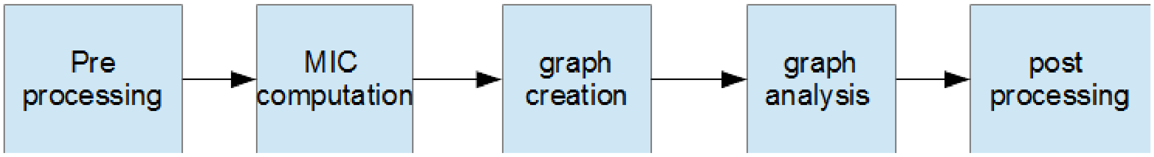


Figure 22 – Overall Pipeline for Network Analysis of Correlated Mutations.

Preprocessing

Figure 23 depicts the various steps of our preprocessing workflow. After searching for strains for a required dataset in IRD, we downloaded the strains as a single FASTA file. Since this FASTA file contains sequences from multiple IAV proteins, we split it into 10 separate FASTA files containing protein sequences for HA, NA, M1, M2, NP, PA, NS1, NS2, PB1 & PB2 proteins. We aligned these sequences using the MUSCLE alignment program. We then performed de-duplication of the sequences across strains by concatenating the sequences of all 10 proteins for a given strain and looking for a 100% match. Strains with duplicate sequences are discarded. We generated 10 FASTA files with sequences for each of the 10 proteins after de-duplication.

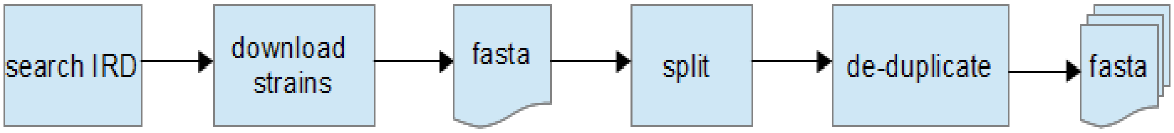


Figure 23 – Preprocessing step where strains of IAV are downloaded, de-duplicated and split into individual protein-specific sequence files to enable downstream processing

MIC Computation

In our MIC-Computation pipeline (depicted in Figure 24), we computed pair-wise MIC correlation coefficient between residues in IAV proteins. We have focused our efforts only on correlations between inter-protein correlations and therefore we did not compute correlations between residues within a protein. For a set of 10 proteins, we have a total of 45 combinations of these proteins. This pipeline is invoked once for each combination (examples: [HA, NA], [M1, M2], [NP, NS1], [NA, PB1]) of proteins. There are two steps in this sub-flow that requires elaboration. First, we must matched the strains from the two FASTA files and create matrices that can be used for pair-wise MIC computation. Second is an important step where the sequences in the matrices must be converted to a Boolean [0,1] format based on the following convention: For each position *i* of a given protein (in the two proteins), the type of amino acid *s* of the multiple sequence alignment (MSA) is represented by a binary variable $x_i(s)$ where

$x_i(s)$ = 1 if the amino-acid is the most frequent amino acid at this position within the MSA

$x_i(s)$ = 0 if it is another amino acid.

x_i = 0 for all sequences in the MSA if number of gaps at position i exceeds 10%

Before we computed MIC, we also discarded all positions with zero variance to improve the speed of computation. We computed MIC between 2 residue positions using functions in minepy library [99]. We created two sets of csv files, the first set consisted of data for all MIC scores greater than 0.1 and the second set consisted of data for all MIC scores greater than 0.5.

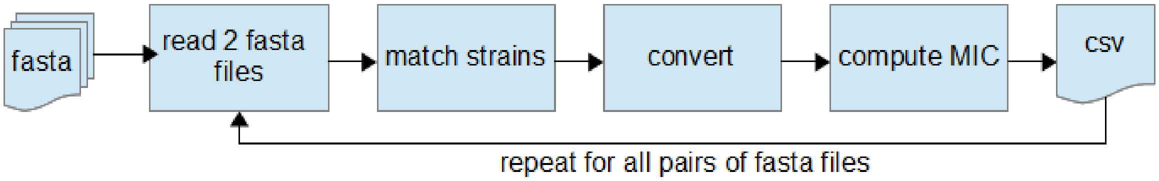


Figure 24 - MIC Computation pipeline where pairwise residue MIC computation is performed between residues in two proteins

Graph Creation

Graph Creation sub-flow is responsible for creating correlated mutation graphs for each of our datasets. Our approach for the design of this sub-flow was to create graphs that would provide the best analysis capabilities, and we decided to create these graphs using two different methods and tooling to realize that objective. We created property graphs in a Neo4j graph database as our first method since this approach gave us a great deal of (querying) flexibility and ability to scale (for additional datasets). For our second approach, we created 'graphml' files that represented our mutation graphs and this allowed us to exploit features in Gephi for visualization and analysis of global characteristics. Figure 25 depicts this workflow with two separate branches for our two approaches. We stored all nodes and edges where MIC > 0.1 in neo4j database since we were interested in understanding the overall node and edge counts for different cutoffs. We used a MIC value of 0.5 as the notional threshold where the correlation is significant, and hence created "graphml" files based on csv files generated using a 0.5 threshold for MIC.

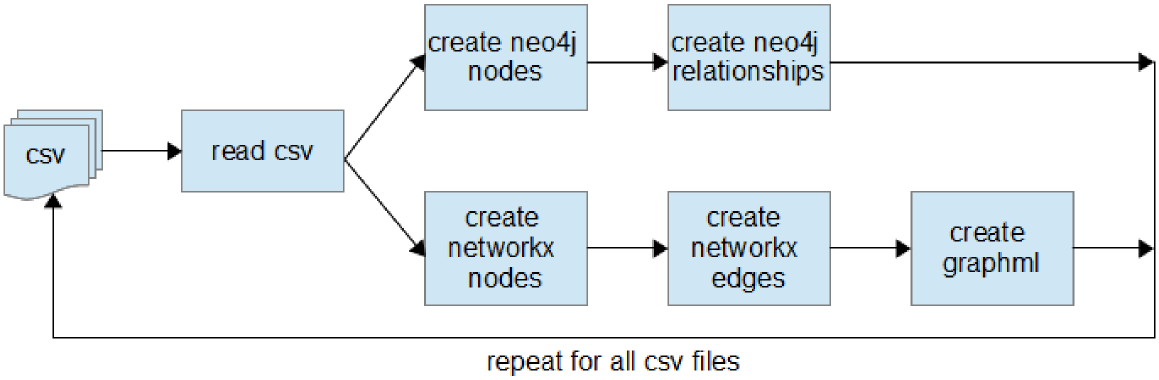


Figure 25 - Graph Creation pipeline where graphs are created based on inter-residue correlated mutations

Neo4j Database Design

We conceived a simple graph database design to meet our requirements. To store several mutation graphs in a single instance of Neo4j, we used the name of the dataset as one of our node labels to give us the required separation of context. We have also created the following uniqueness constraint in our database to prevent duplicate nodes.

create constraint on (p:P) assert p.unique_id is unique

where P = a protein from {HA, NA, M1, M2, NP, PA, PB1, PB2, NS1, NS2}

and unique_id = concatenation of dataset_name + protein +residue_number

Table 6 provides a listing of labels and properties in our Neo4j graph database. We have purposefully added some redundancy into the label names and property names to provide us maximum flexibility and query optimization, since labels can provide us a higher-level separation (based on dataset name and protein) while property labels can provide us greater querying capabilities.

Table 6 - Neo4j labels and properties

Node Properties	protein (2 letter name of protein) aa (3 letter amino acid code) residue_number (residue number in protein sequence) dataset (name of dataset) unique_id (concatenation of dataset_name + protein +residue_number, used for uniqueness constraint)
Edge Properties	mic (MIC score)
Relationship Types	MUTATES_WITH (relationship between 2 nodes)
Node Lables	dataset (name of dataset) protein (2 letter name of protein)

Graph Analysis

Our computational pipeline (depicted in Figure 26) for ‘graph analysis’ is comprised of two methods.

1. Creation of a web application that allows users to
   a. Search connections of a specific node

   b. View plots depicting number of nodes & edges for different threshold values

   c. View top nodes, edges and triplets for each dataset

   d. View global properties of graphs generated with a MIC filter > 0.5 including degree distribution, clustering coefficient, edge density and average degree

   e. View interactive network visualizations for correlated mutation graphs (generated with a MIC filter > 0.5)

2. Use of Gephi visualization and exploration platform to perform analysis of correlated mutation graphs. We have included results of any analysis that we did using Gephi (including visualizations, images, other data points) as part of our web application.

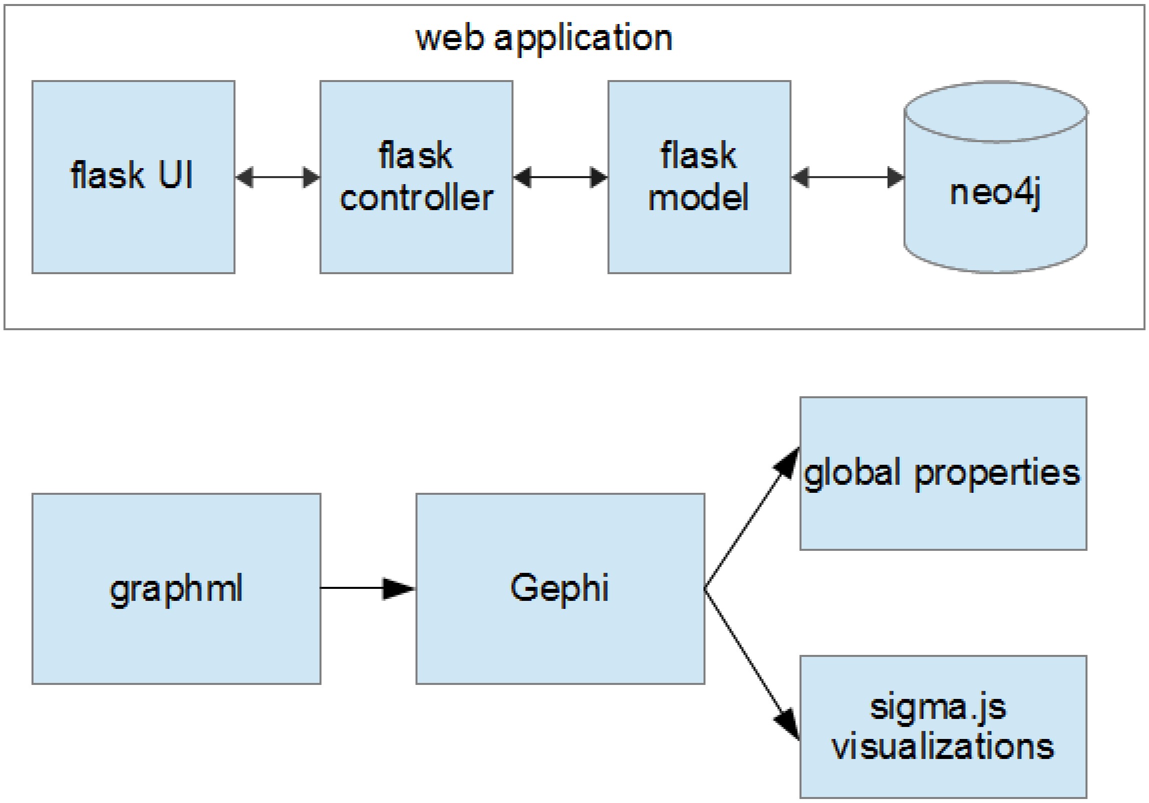


Figure 26 - Graph Analysis pipeline for analysis of correlated mutation graphs

Additional Post-processing

In addition to conducting analysis of networks generated for different datasets to determine structural and topological properties, we performed post-processing steps (Figure 27) to answer specific questions. First, we conducted entropy analysis to understand the amount of variation in different proteins in all datasets and thereby identify potential associations between entropy and MIC correlations. Second, we were interested to know if most of the in-network residues are on the surface of a protein with higher solvent accessibility values and we conducted solvent accessibility analysis to answer this question. Third, we calculated residue cooccurrence counts for pairs to get a better insight into the underlying mechanics of MIC algorithm for our use case. Finally, we created 'protein networks' to provide a macro picture and depict the extent of correlations between proteins.

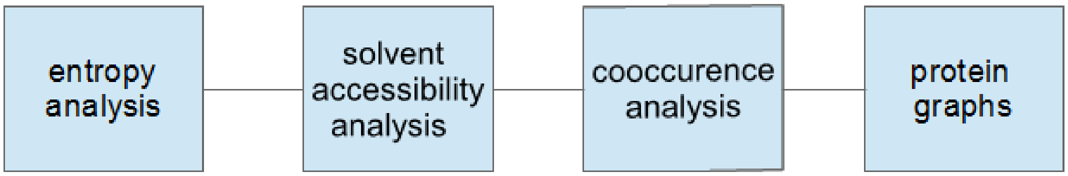


Figure 27 - Additional post processing steps that we conducted to provide additional context to our results

CHAPTER 8 - RESULTS

In this chapter, we present out results based on comprehensive network analysis of correlated mutations on multiple IAV datasets. We start with plots of node counts and edge counts for different MIC threshold values, since this association can provide a general understanding of the degree of covariance. We also present MIC distribution plots since the network structure and dynamics is directly related to these distributions. Based on our general assumption that correlations that exceed a MIC value of 0.5 should be significant or near-significant, we create degree distribution and clustering coefficient distributions for all 10 datasets previously listed in the section on Chapter 5 - DataSets. Given our interest to understand the potential relationship between entropy and correlated mutation networks, we include details of average entropies for each of the proteins in all datasets. We create schematics of protein-level graphs for correlated mutation networks @ 0.5 MIC, to provide more obvious protein-level insights. We delve into the details of residue combination counts of highly significant, significant and not significant edges to gain a better perspective into the underlying mutational patterns and to ensure that MIC is a reliable statistic.

Web Application

We have created a web application to share the results of our work with the broader community. This web application can be accessed at http://omics.gmu.edu:5000. While majority of the images and visualizations depicted in this application (node counts, edge counts, degree distributions and macro plots) are pre-computed and stored, the ‘top N’ and ‘search’ functionalities provide real-time dynamic results based on querying of neo4j database.

Network Visualizations

We created visualizations of correlated mutation networks at 0.5 threshold. Interactive version of these visualizations can be viewed at http://omics.gmu.edu:5000/dviz. We have also included a static version of these network diagrams for human H3N2, swine H3N2, human H1N1, swine H1N1 and avian H5 datasets in Figure 28, Figure 29, Figure 30, Figure 31 and Figure 32.

These pictures clearly elucidate that there are significant differences in structural topologies of these networks. The density of nodes in human H1N1 network is significantly higher that swine H1N1 and the same comparison holds true for human H2N2 network over swine H3N2 network. The avian H5 network is dominated by residues from NA (purple) while the swine H1N1 network is characterized by four distinct clusters.

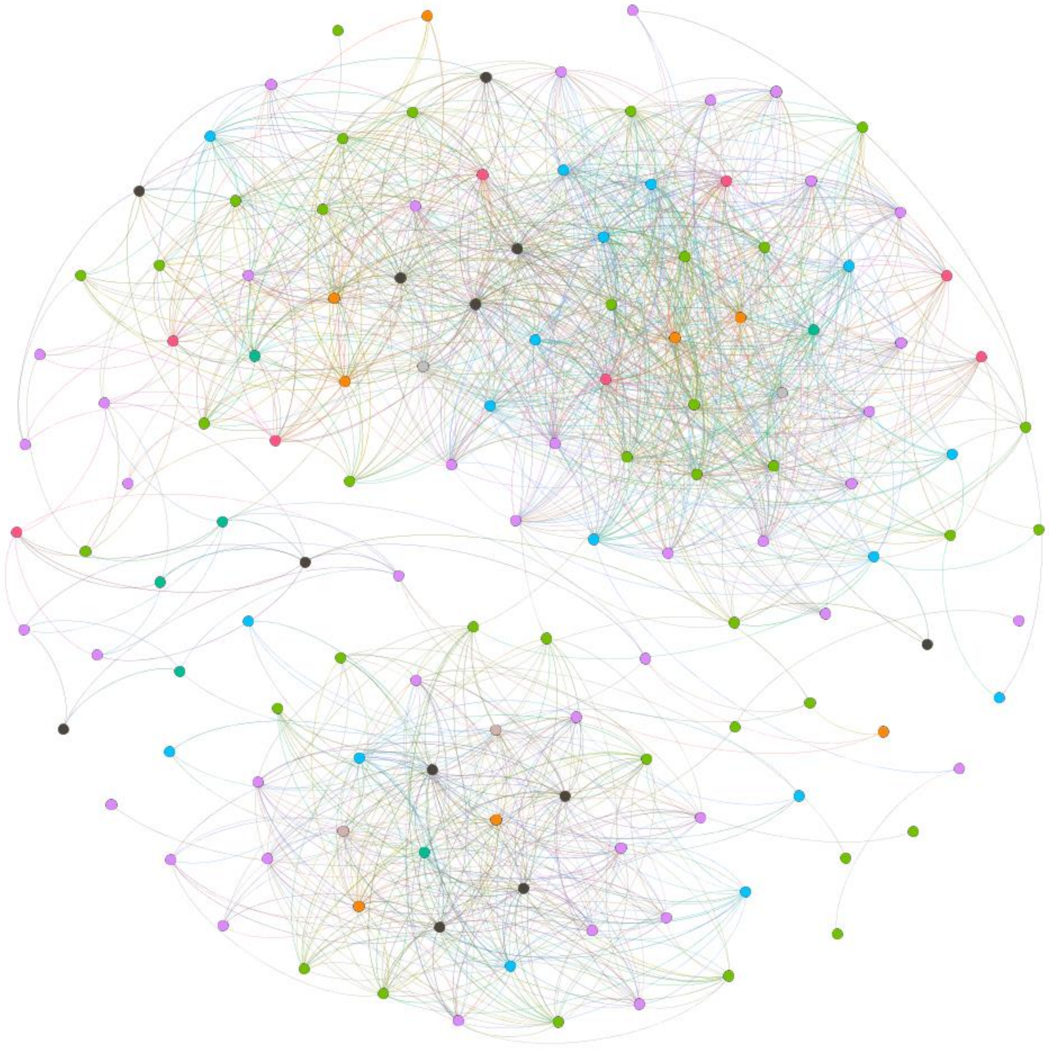


Figure 28 - Visualization of human H3N2 correlated mutation network at 0.5 threshold

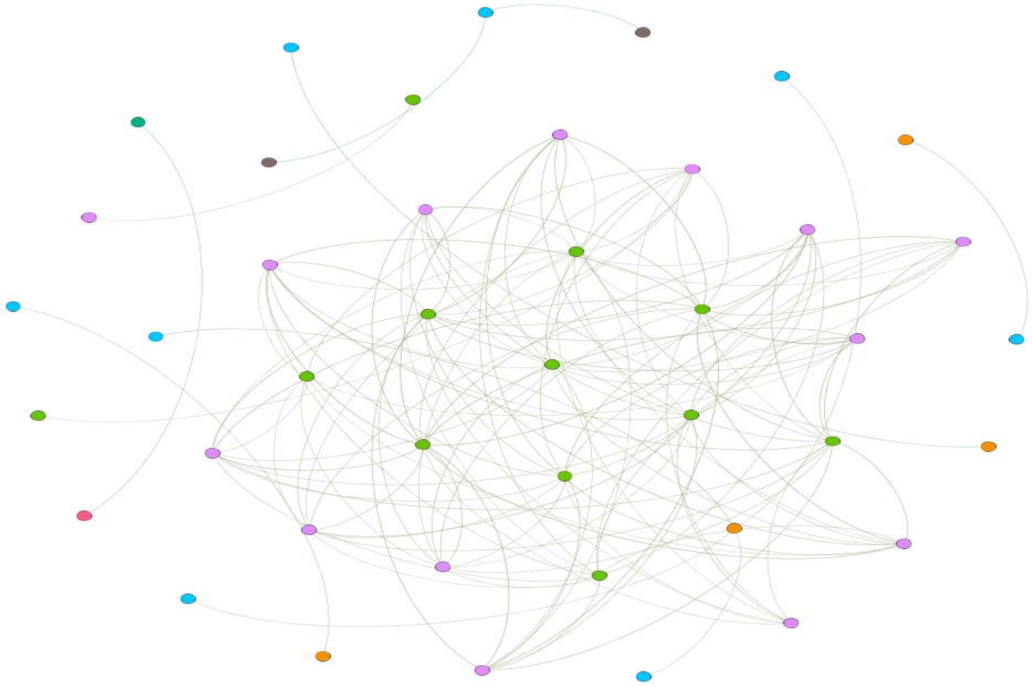


Figure 29 - Visualization of swine H3N2 network at 0.5 threshold

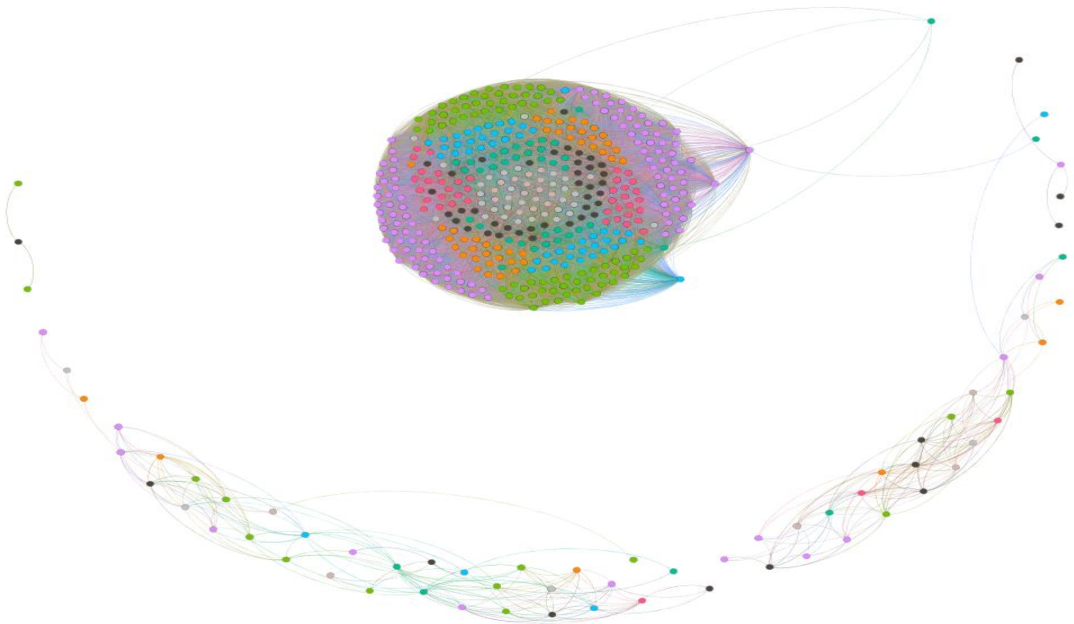


Figure 30 - Visualization of human H1N1 network at 0.5 threshold

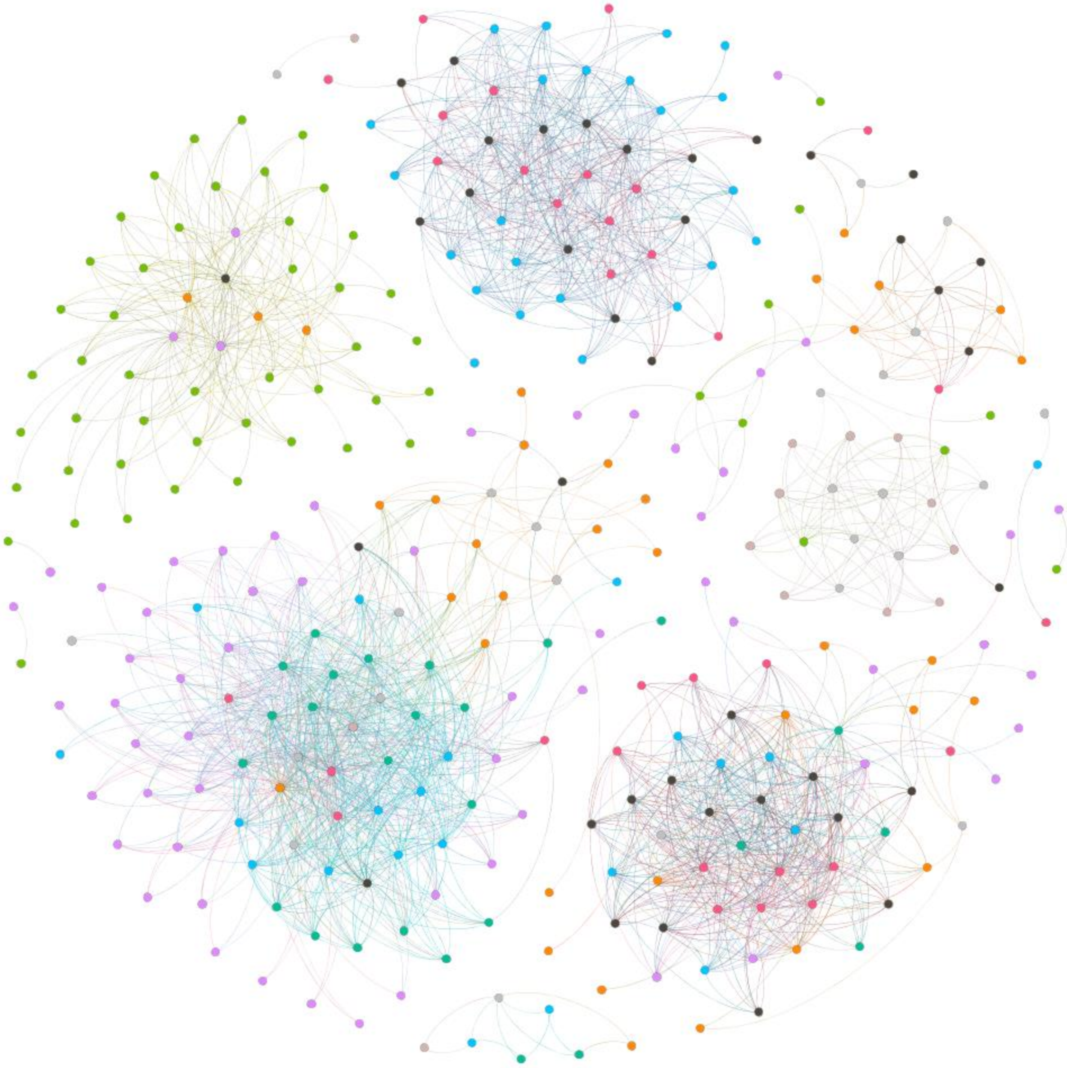


Figure 31 - Visualization of swine H1N1 network at 0.5 threshold

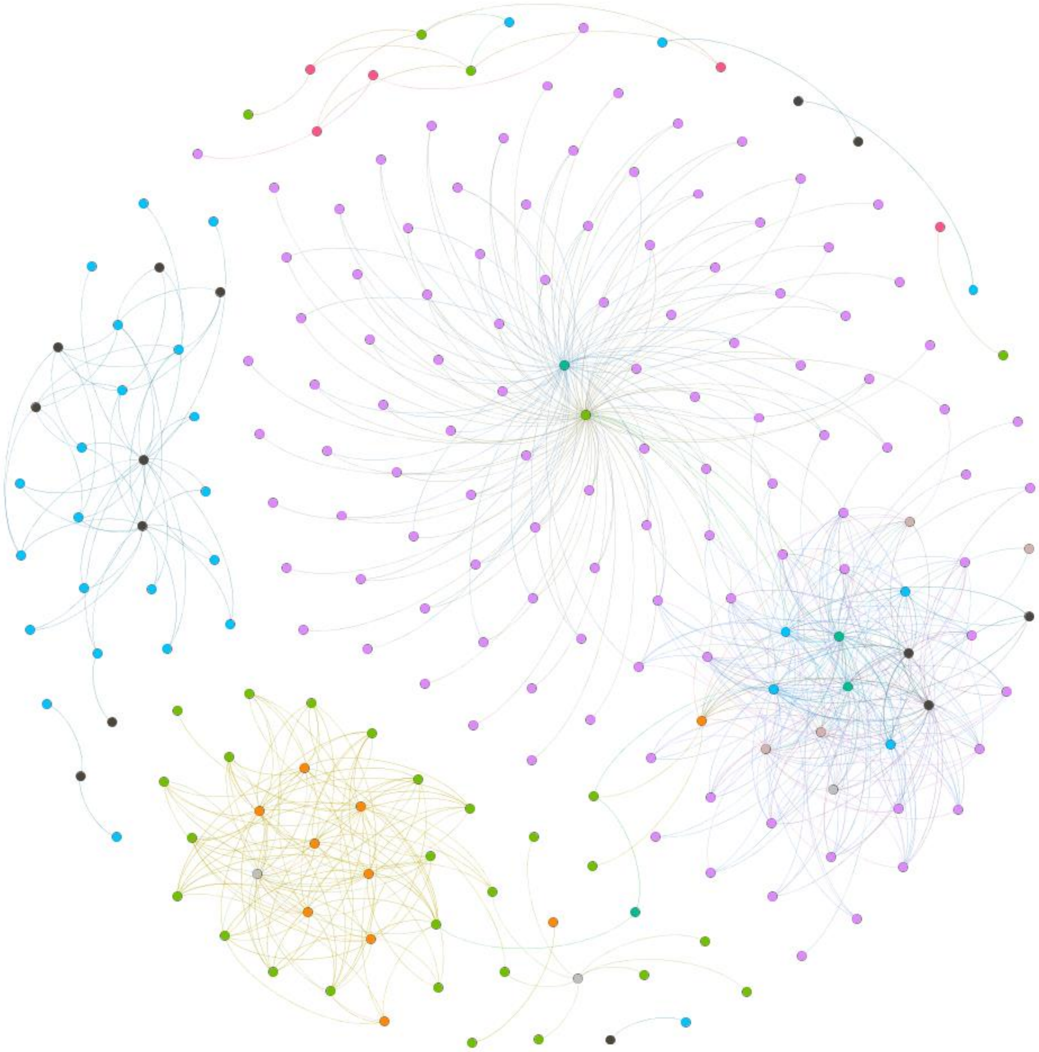


Figure 32 - Visualization of avian H5 network at 0.5 threshold

Node Counts

IAV strains that we downloaded from IRD consisted of a total of 4499 residues. We computed the total number of in-network residues for different MIC threshold values and noticed a wide variance based on the dataset. The number of nodes gradually decreases as MIC threshold increases. Several interesting observations can be made based on plots in Figure 33, Figure 34, Figure 35, Figure 36 and Figure 37.

1. Human H1N1 vs Swine H1N1 (Figure 33) - The number of in-network residues for human H1N1 IAV correlated mutation network tends to be stable till a MIC threshold of ~0.6 after which it starts to decline while the number of in-network residues for swine H1N1 network gradually decreases for increasing values of MIC threshold. It should be noted here that the number of residues in human H1N1 network with MIC correlations in (0.1, 0.5) range is very low implying that we do not see new nodes joining the network for lower MIC values. While we see a slightly higher number of nodes in swine H1N1 at low threshold values in (0.1, 0.4) range, these numbers gradually decrease and the number of nodes with at least one significant mutation is lower compared to human H1N1 network. Both networks have similar node count for MIC values > 0.8.
2. Human H3N2 vs Swine H3N2 (Figure 34) - We see a high number of in-network nodes (~650) for swine H3N2 strains at a low threshold value of 0.1 and we see a steeper decrease for increasing MIC threshold values and the number comes down to less than 50 for MIC threshold values in the significance region (>0.5). In other words, 98% of the residues in swine H3N2 strains do not have any significant correlated mutations. The human H3N2 network starts with a lower number of nodes (~350) at 0.1 MIC threshold and we see a much more gradual decrease in the number of nodes.

We also observe a roughly overlapping tail for MIC > 0.88 in both these plots.

3. There are approximately 25 nodes in the H7N9 network in the significant zone (MIC > 0.5) while there are approximately 200 nodes in the significant zone for avian H5 network.
4. The results that we observe in Figure 37 for the "Human All" network are generally suggesting that there will be higher degree of covariation when we mix strains belonging to different sub-types.

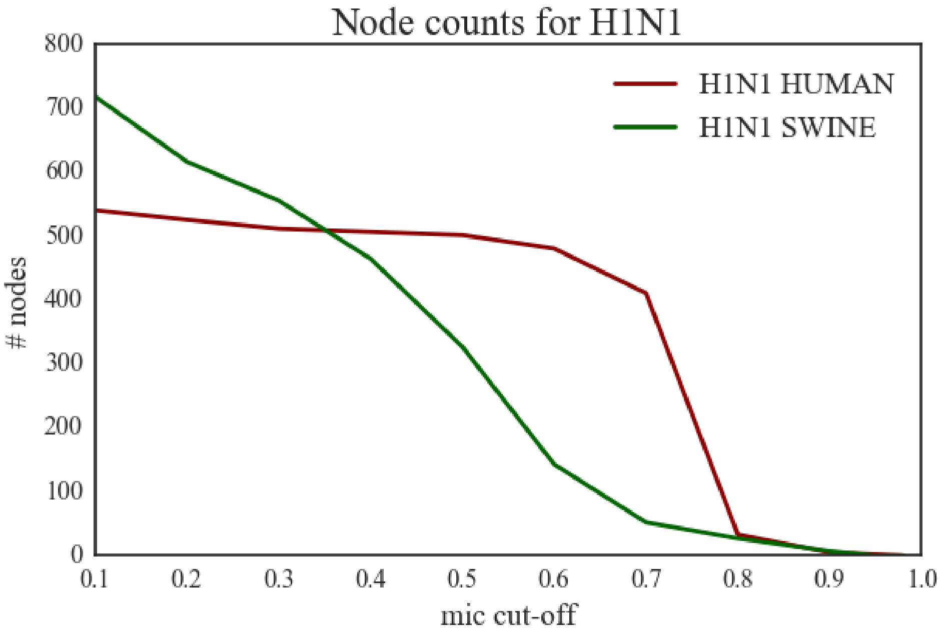


Figure 33 - Node counts for H1N1 datasets.

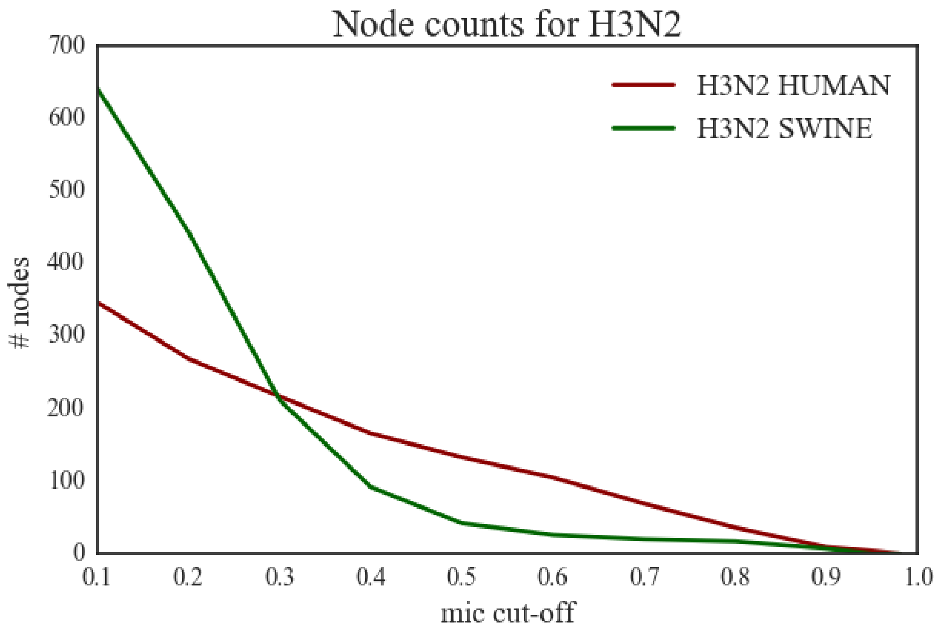


Figure 34 - Node counts for H3N2 datasets

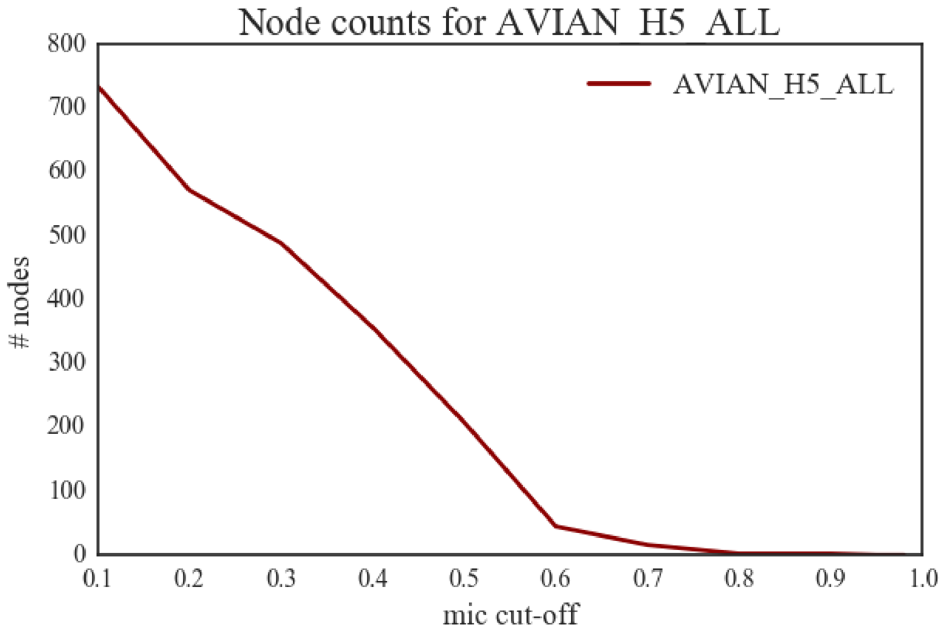


Figure 35 - Node counts for AVIAN H5 dataset

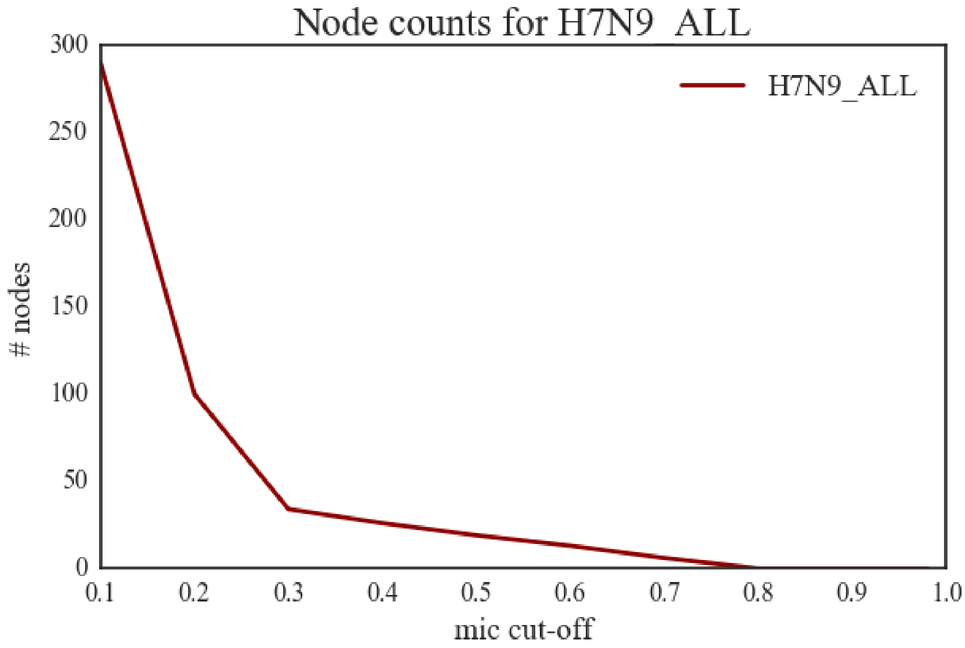


Figure 36 - Node counts for H7N9 dataset

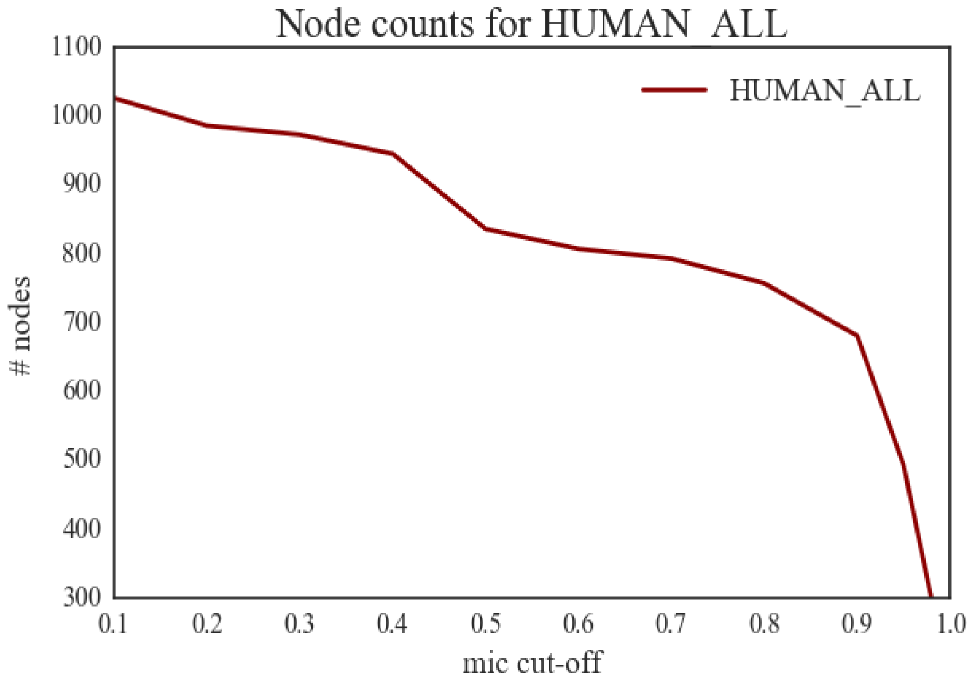


Figure 37 - Node counts for HUMAN_ALL dataset

Edge Counts

The theoretical upper limit (max) for the total number of inter-protein edges in IAV strains approximates to 8.8 million. We computed the total number of edges in the network for different MIC threshold values. Plots for edge counts are depicted in Figure 38, Figure 39, Figure 40, Figure 41 and Figure 42.

1. If we exclude the ‘Human All’ network, we can make a general conclusion that less than 2% of edges in IAV have correlated mutations (MIC > 0.1) and a smaller fraction of edges have significant correlations.
2. There is a wide variance in edge counts based on the dataset.

3. H1N1 vs. H3N2 – The number of edges in H3N2 networks is very small compared to edges in H1N1 networks.
4. Human H1N1 vs. Swine H1N1 – There is a significant difference between the number of edge counts for MIC > 0.5 in these two networks. There are 79524 significant edges in human H1N1 compared to a much smaller number (501) in swine H1N1.
5. Human H3N2 vs. Swine H3N2 – swine H3N2 network contains only 132 edges in the significant zone while the human H3N2 network contains 1378 significant edges. There is a steep decline in the number of edges (from 290000 to 5000) in swine H3N2 as the MIC threshold changes from 0.1 to 0.2.
6. We see more than 40000 edges in avian H5 for 0.1 MIC threshold but only 647 of these edges have MIC values greater than 0.5.
7. The H7N9 network is characterized by a very small number of edges. This network contains only 37 edges with MIC values greater than 0.5.
8. Like our observation with ‘node counts’, the number of edges in ‘Human All’ network is significantly higher compared to all other datasets.

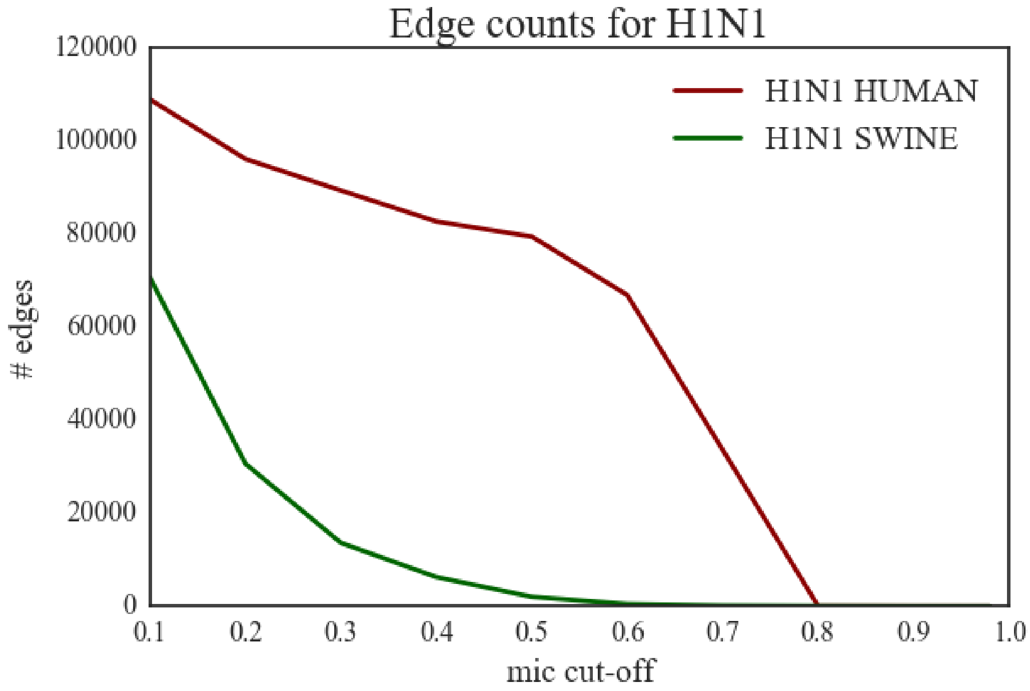


Figure 38 - Edge counts for H1N1 datasets

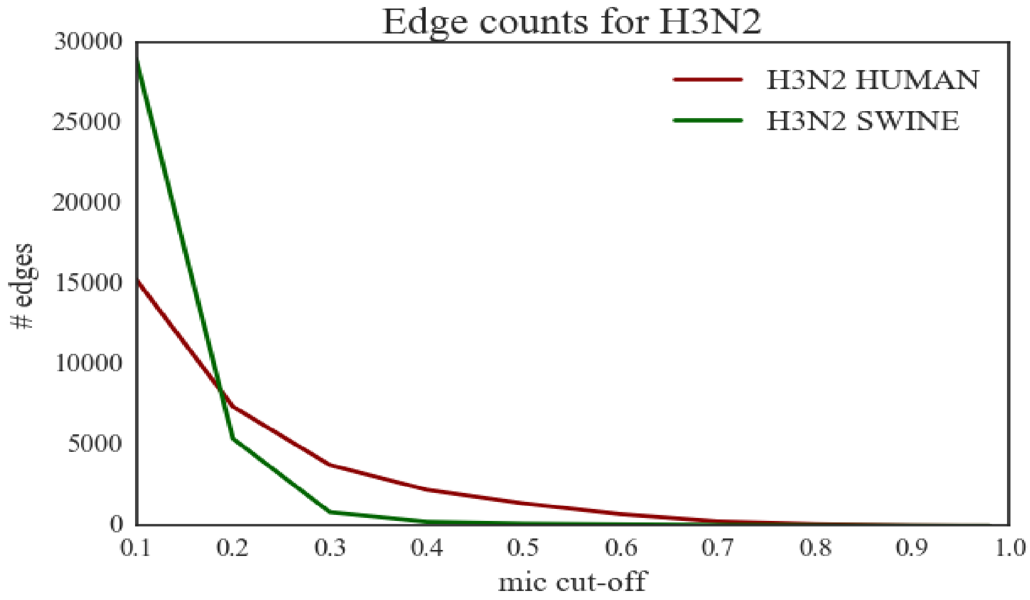


Figure 39 - Edge counts for H3N2 datasets

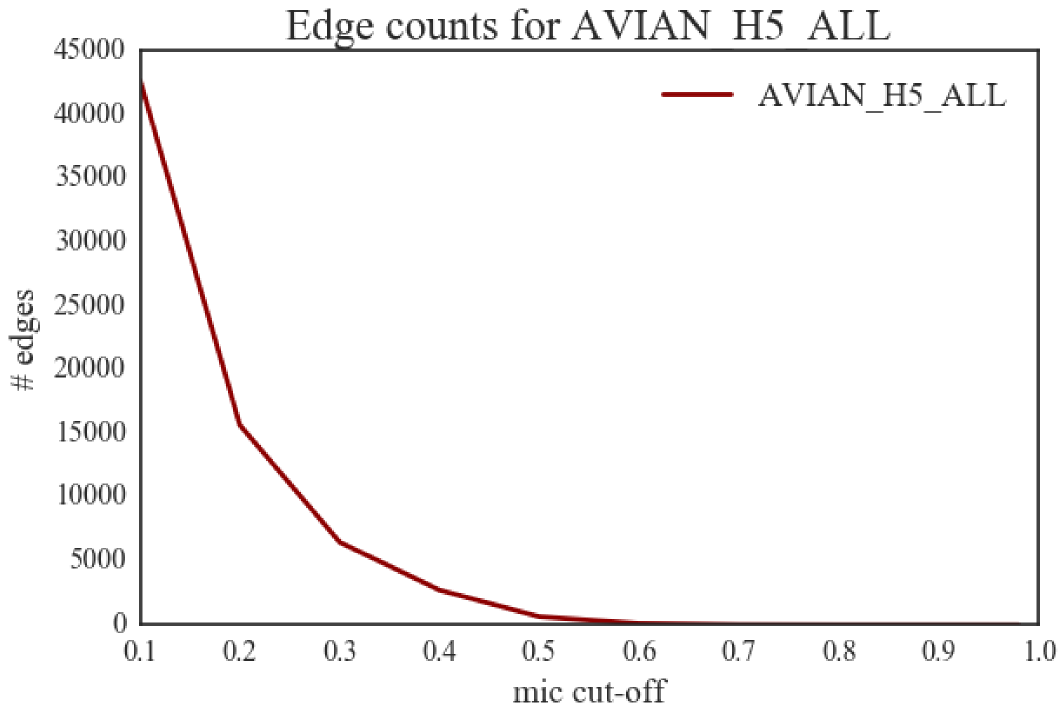


Figure 40 - Edge counts for AVIAN H5 dataset

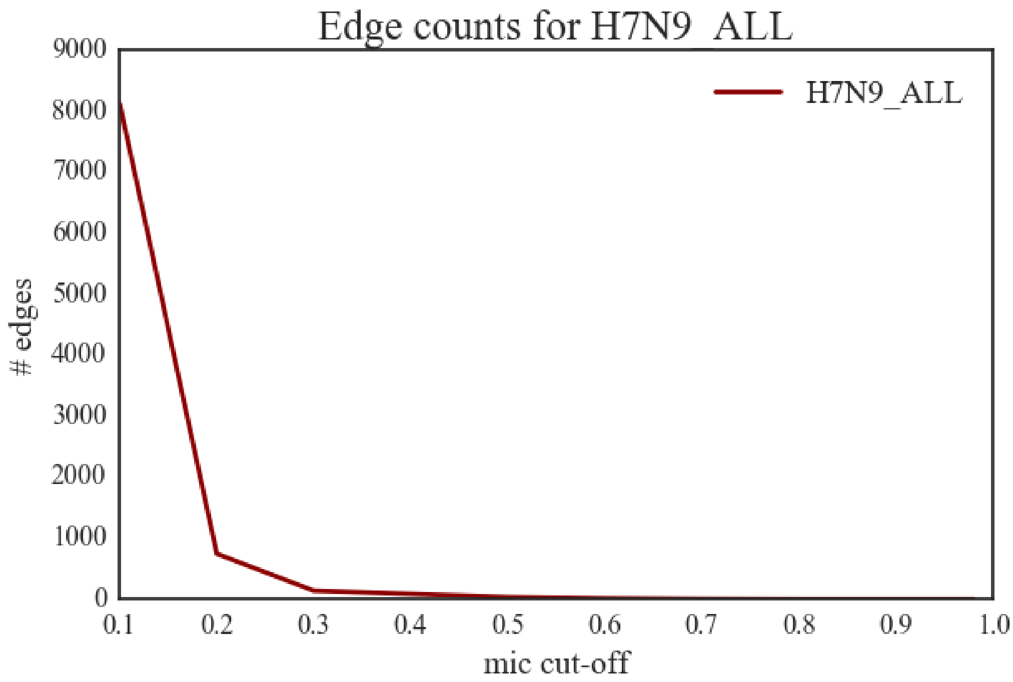


Figure 41 - Edge counts for H7N9 dataset

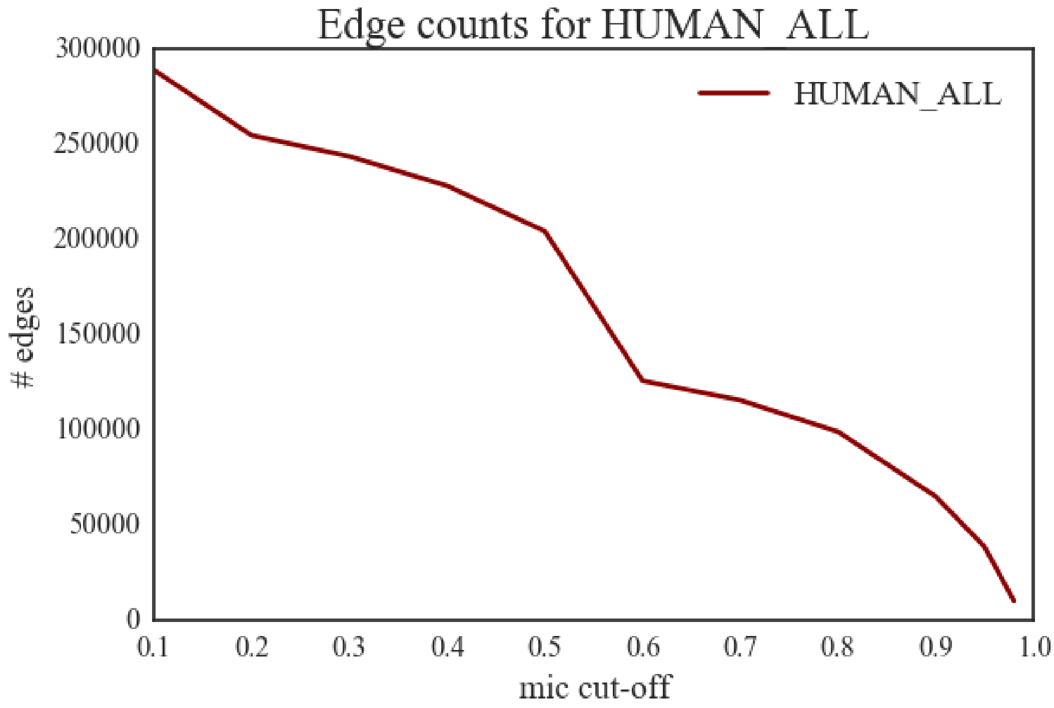


Figure 42 - Edge counts for HUMAN ALL dataset

MIC Distribution

The MIC distribution plots depicted in Figure 43, Figure 44, Figure 45, Figure 46, Figure 47, Figure 48 and Figure 49 provide us additional insight into the extent of covariance between residues in IAV proteins and provide us appropriate reasoning behind the change in node and edge counts as we increase the MIC threshold. The MIC histograms for Swine H1N1, Swine H3N2, Human H3N2 Avian H5 and H7N9 confirm to a 'power law' model suggesting that there is a high degree of concentration of residues with low MIC values and the correlation decreases rapidly for higher MIC threshold values. Other histograms (corresponding to Human H1N1 and Human-All datasets) do not have a specific pattern associated with them, with the 'Human-All' distribution diverging the most from a ‘power law’ model with a relatively high concentration of residues in the [0.9, 1] range.

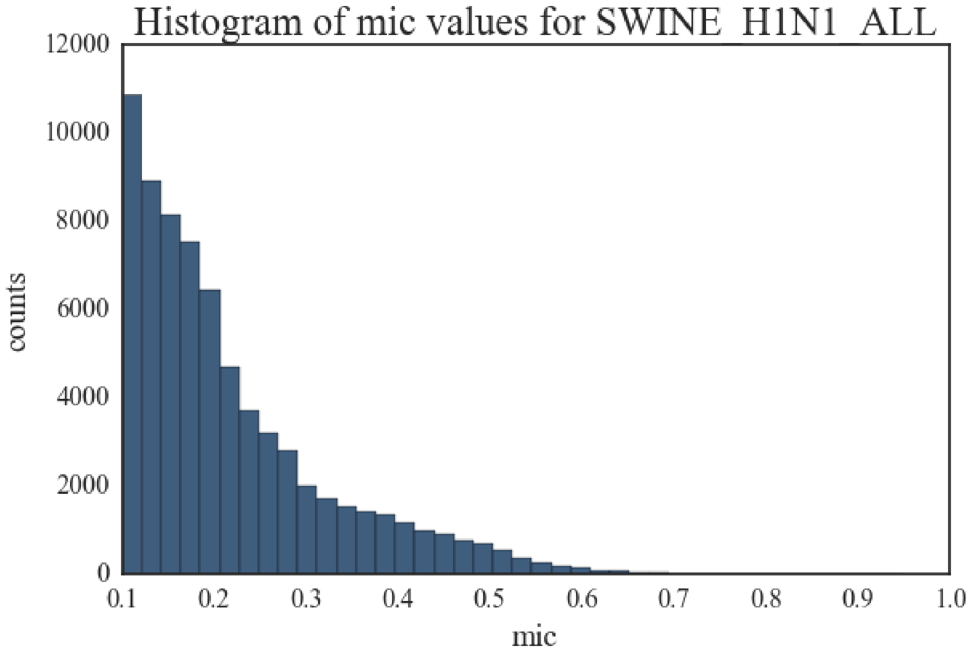


Figure 43 - MIC distribution for SWINE H1N1 dataset

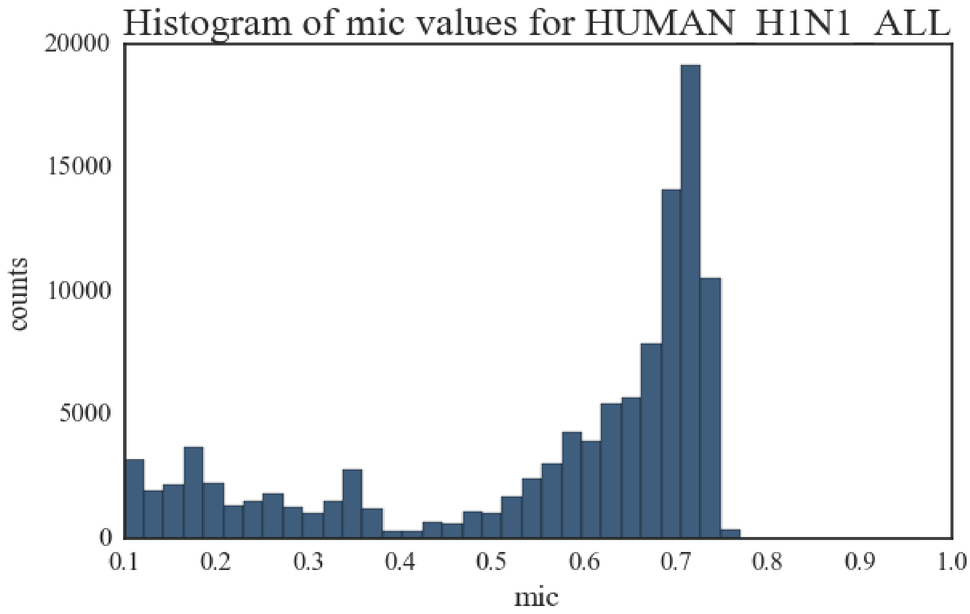


Figure 44 - MIC distribution for HUMAN H1N1 Dataset

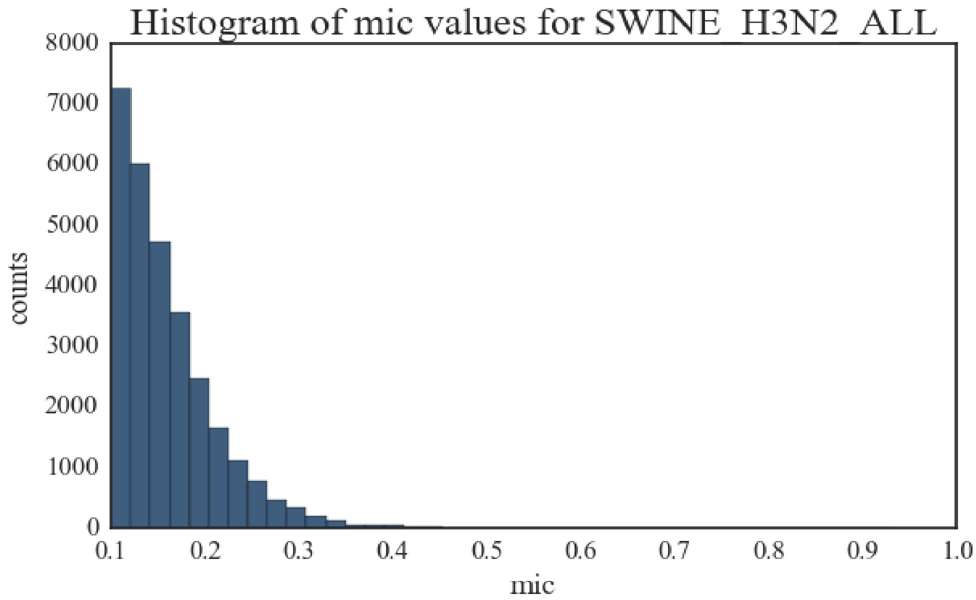


Figure 45 - MIC distribution for SWINE H3N2 dataset

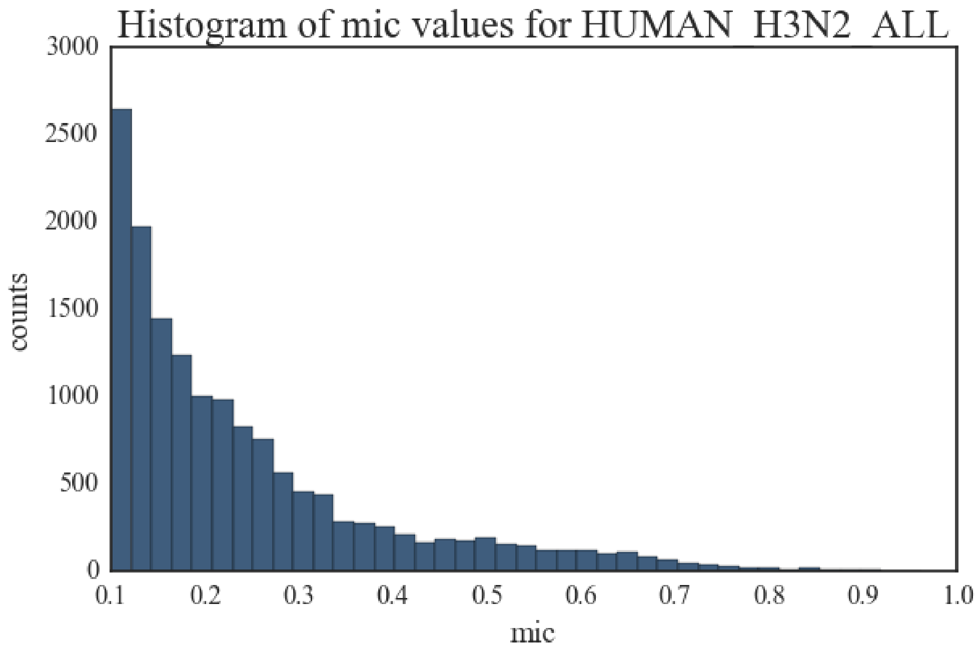


Figure 46 - MIC distribution for HUMAN H3N2 Dataset

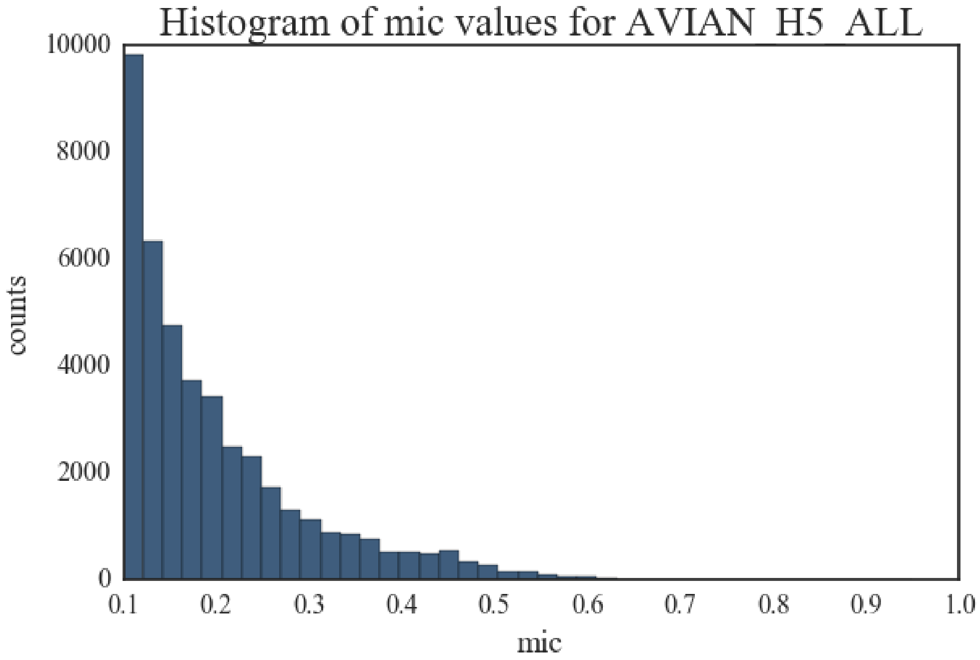


Figure 47 - MIC distribution for AVIAN H5 dataset

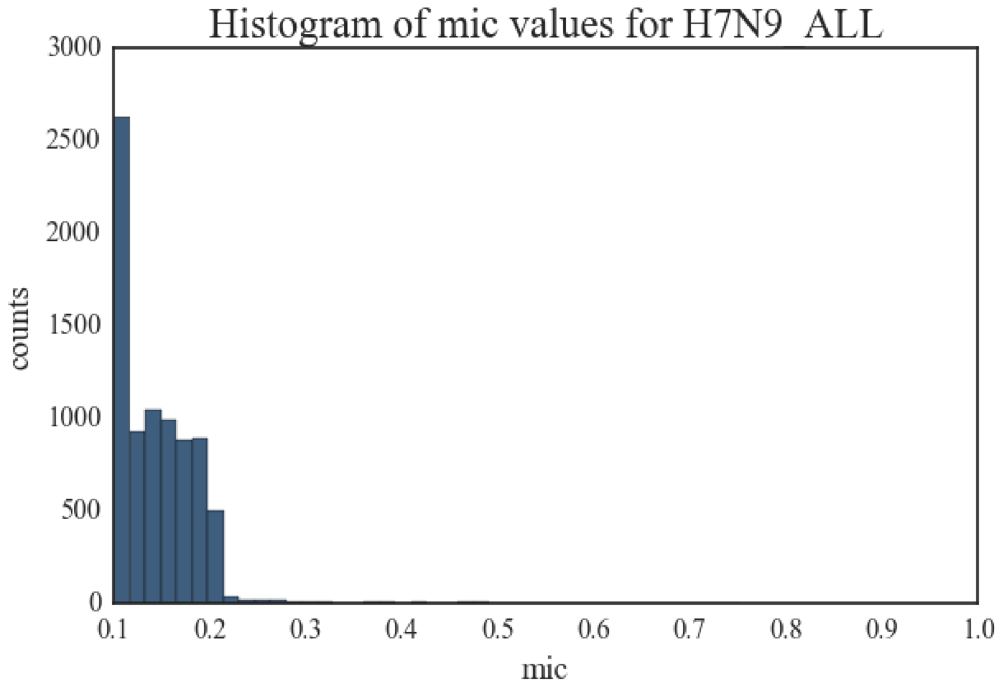


Figure 48 - MIC distribution for H7N9 dataset

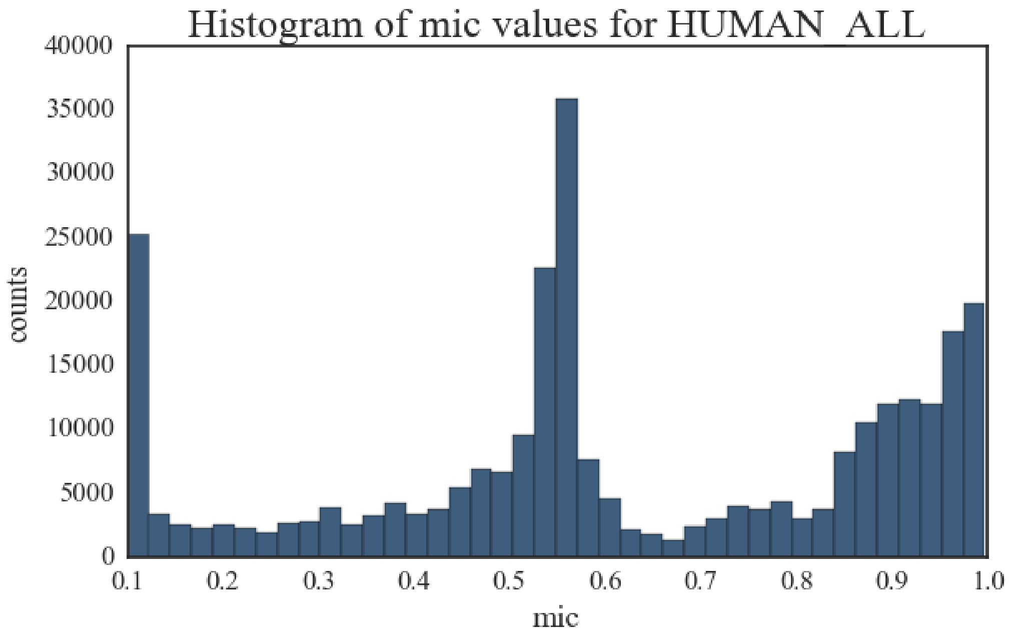


Figure 49 - MIC distribution for HUMAN_ALL dataset

Edge Densities

We have made an empirical decision that MIC values greater than 0.5 indicate significant mutations and hence explored the properties of networks at this value. As a first step, we listed the node and edge counts along with edge density for our primary datasets in Table 7 to further elucidate the differences between networks (for MIC > 0.5). This table reaffirms our earlier observation that the "Human All" network @ 0.5 MIC has higher nodes, edges and more importantly this network is characterized by a high edge density. The 'Human H1N1' and 'Avian H5' networks have the highest and lowest edge densities respectively.

Table 7 - Networks @ 0.5

Dataset	#nodes	#edges	edge density

Swine H3N2	42	132	0.15
Human H3N2	133	1378	0.16
Swine H1N1	327	2023	0.04
Human H1N1	501	79524	0.63
Avian H5	209	647	0.03
H7N9	20	37	0.19
Human All	903	238832	0.59

Degree Distribution

The degree of a vertex is the number of edges emanating from it. Degree distribution is an important characteristic of a graph and provides a distribution of the degree of nodes over the entire network. Degree distribution reflects the overall pattern of connections in a dataset. A node with high degree in correlated mutation network implies that the residue has correlated mutations with many residues in the network.

We provide histograms of degree distributions in Figure 50, Figure 52, Figure 51, Figure 53, Figure 54, Figure 55 and Figure 56. There is substantial difference between each of these degree distributions indicating that the overall structure of a correlated mutation network in IAV does not adhere to a single topology. While the distributions for Swine H1N1 and Avian H5 are close to a power-law model, other network distributions are more indicative of random networks. The network that corresponds to “Human All” dataset is highly connected with many nodes with a degree greater than 500. These plots illustrate the complex evolutionary patterns in Influenza and highlight the fact that the overall mutation profile and evolution in IAV strains are sub-type specific.

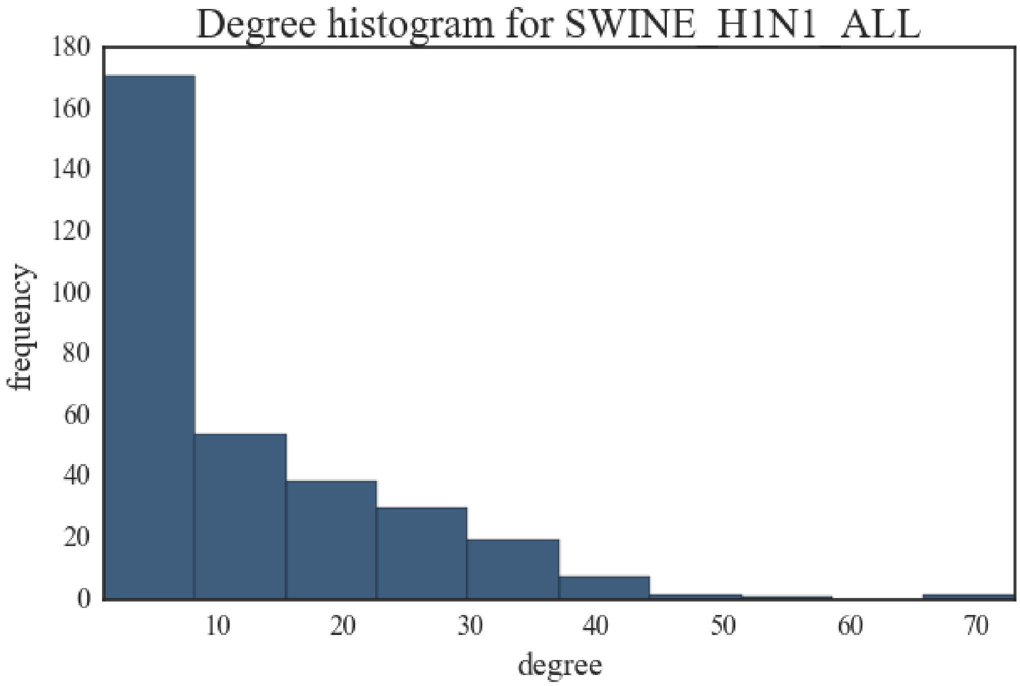


Figure 50 - Degree histogram for Swine H1N1 dataset

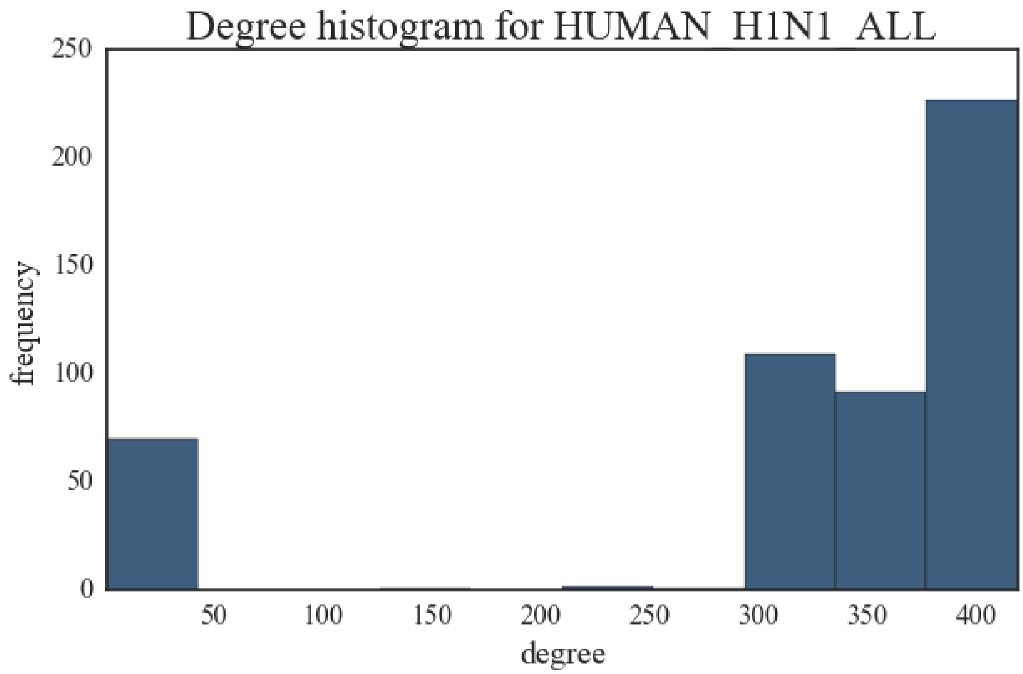


Figure 51 - Degree distribution for Human H1N1 dataset

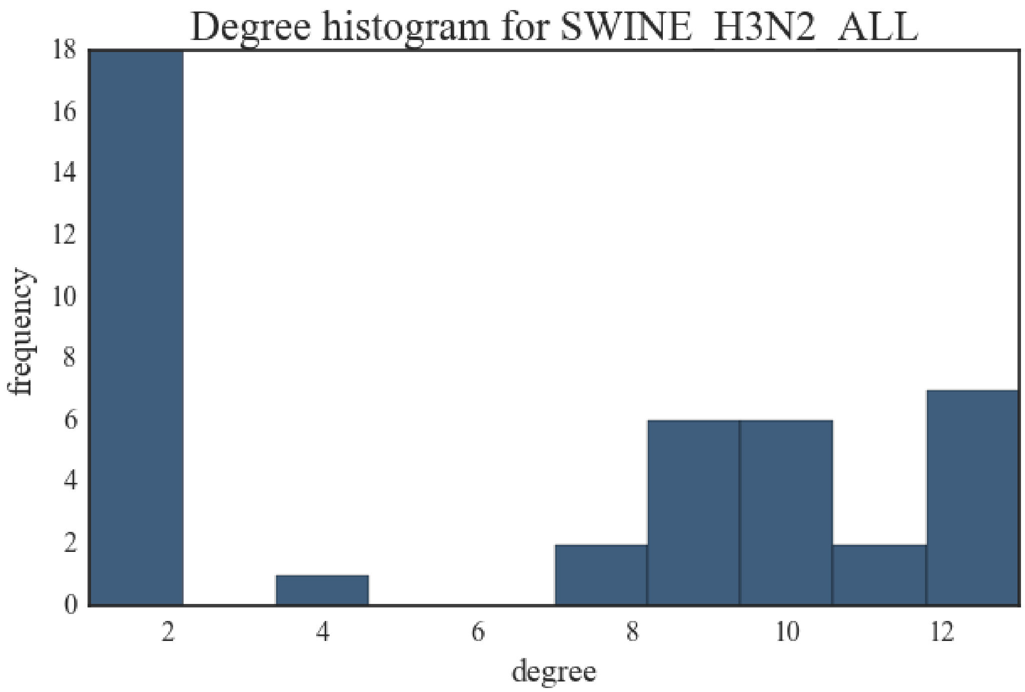


Figure 52 - Degree histogram for Swine H3N2 dataset

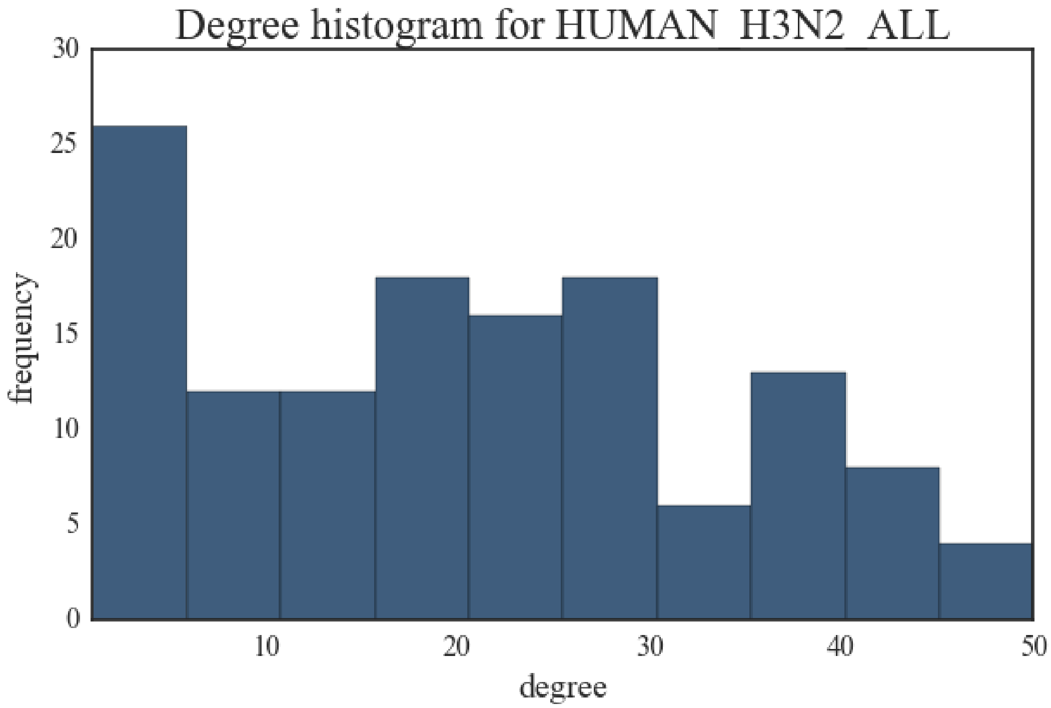


Figure 53 - Degree distribution for Human H3N2 dataset

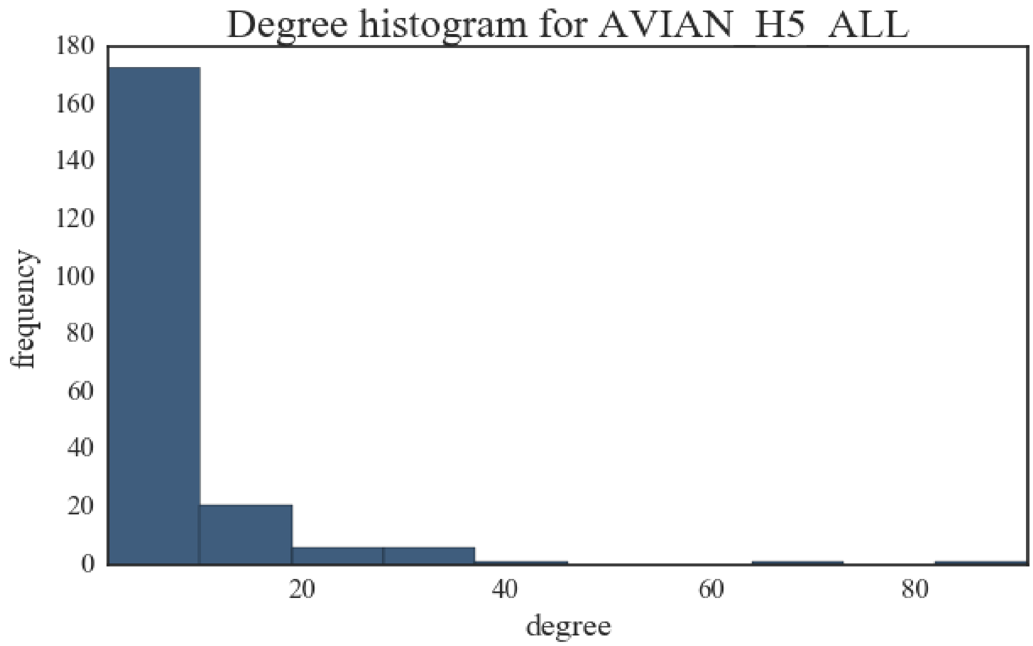


Figure 54 - Degree distribution for Avian H5 dataset

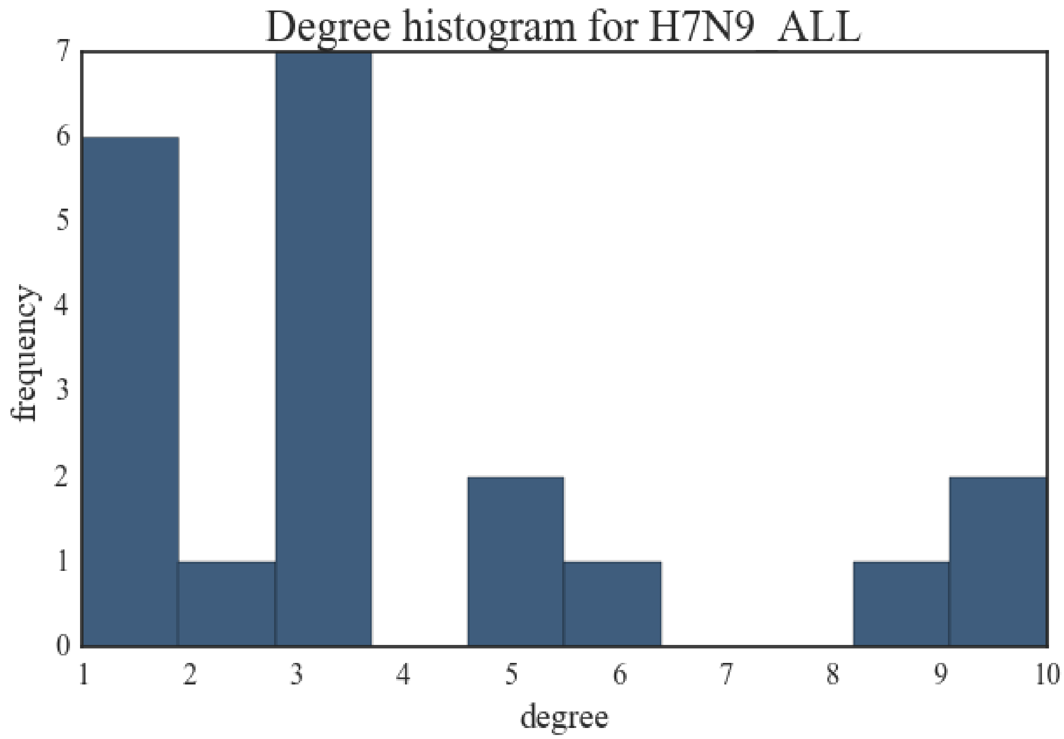


Figure 55 - Degree distribution for H7N9 dataset

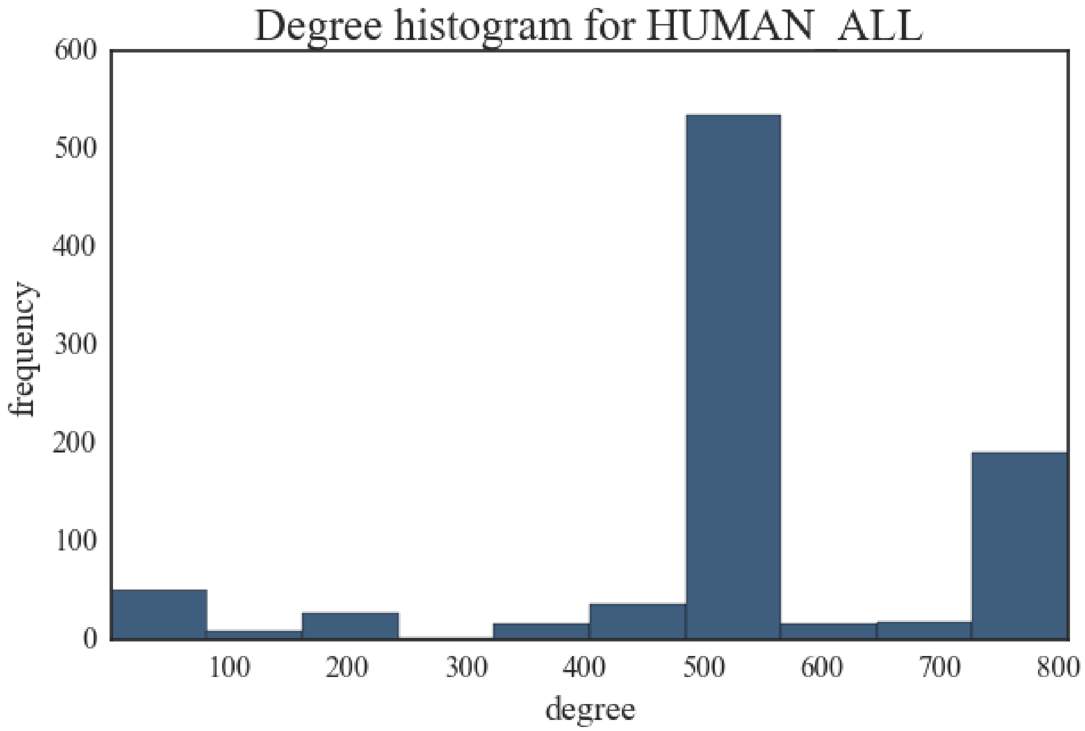


Figure 56 - Degree distribution for Human All dataset

Clustering Coefficient Distributions

The clustering coefficient of a vertex indicates how concentrated the neighborhood of that vertex is. The clustering coefficient is the ratio of the number of actual edges there are between neighbors to the number of potential edges between neighbors (all possible edges between the vertices). A node with high clustering coefficient in correlated mutation network implies that the residue is part of a neighborhood where residues are covarying with high probability.

We provide histograms of local clustering coefficient distributions in Figure 58, Figure 59, Figure 60, Figure 61, Figure 62, Figure 63 and Figure 64. Figure 57 illustrates the difference in average clustering coefficient between datasets.

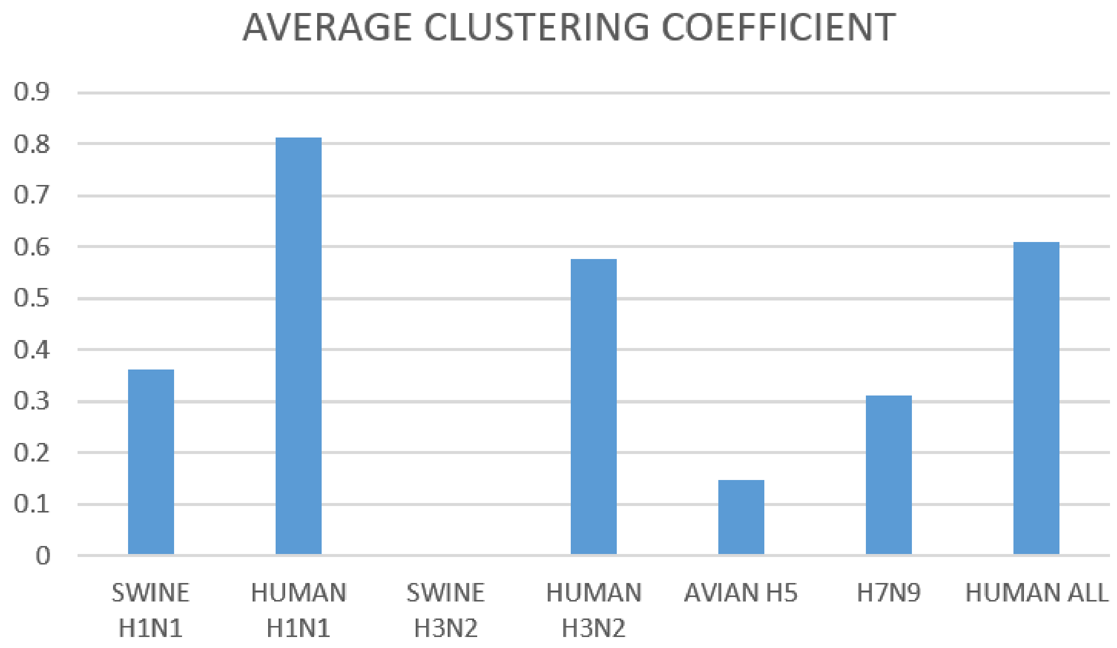


Figure 57 - Average Clustering Coefficient

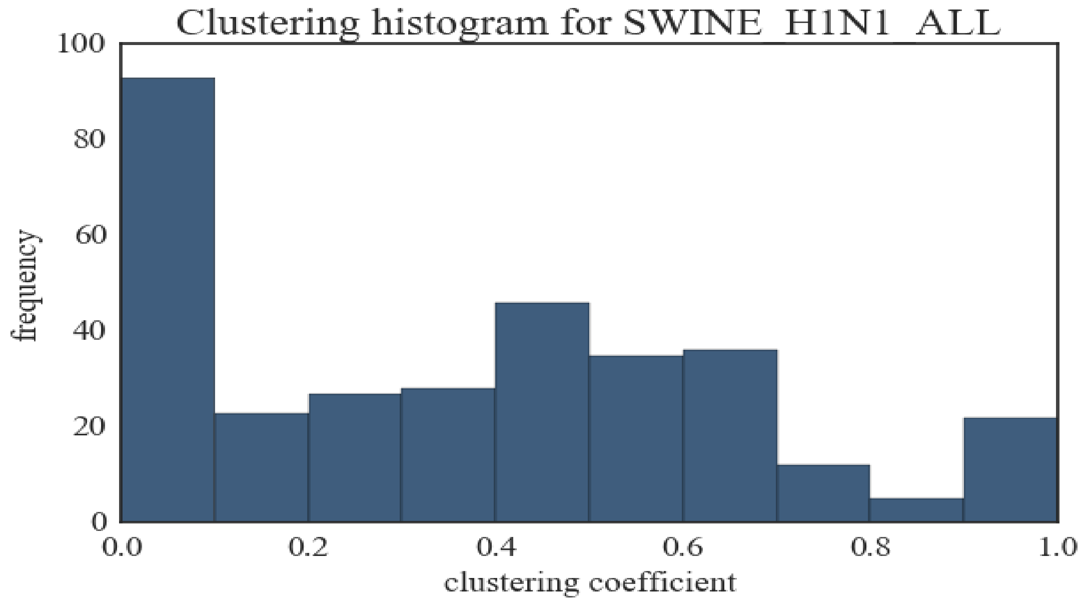


Figure 58 - Clustering coefficient histogram for Swine H1N1 dataset

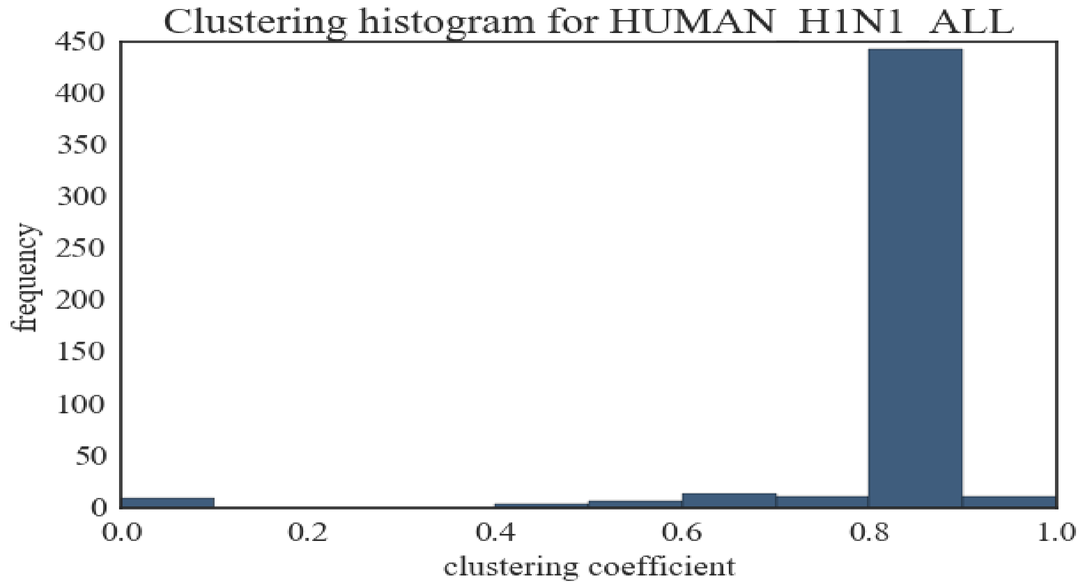


Figure 59 - Clustering coefficient histogram for Human H1N1 dataset

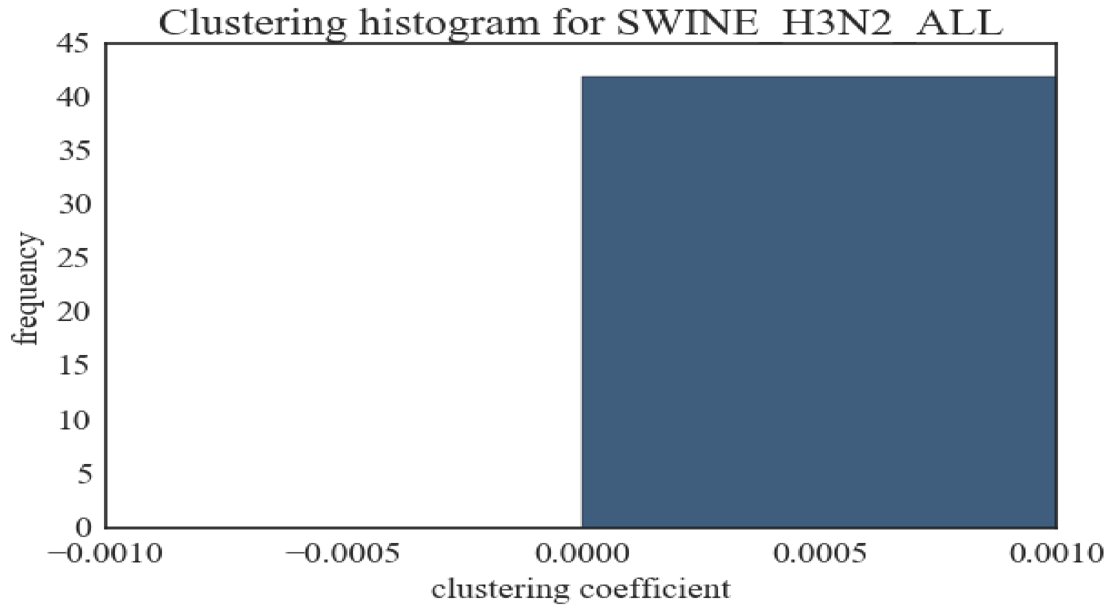


Figure 60 - Clustering coefficient histogram for Swine H3N2 dataset

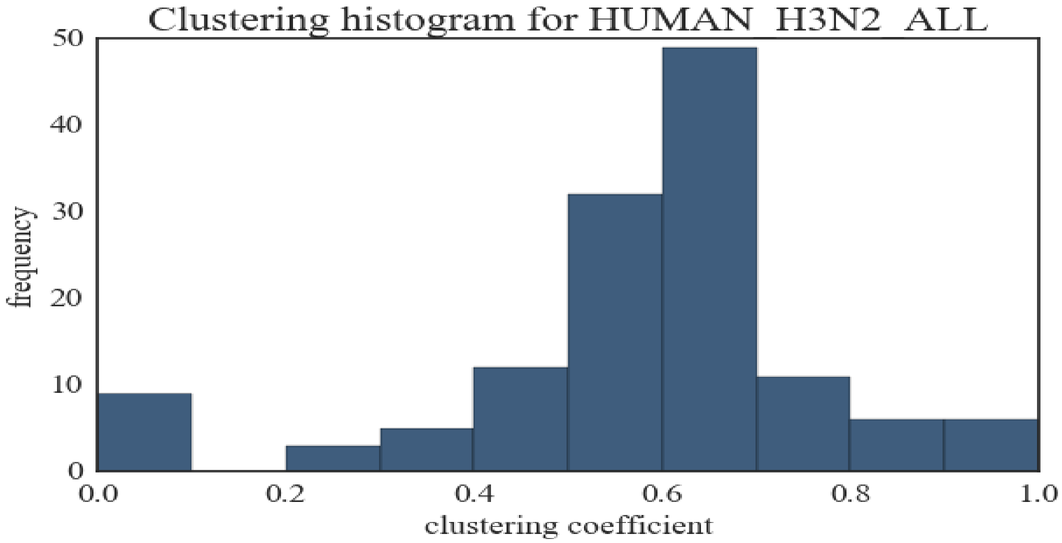


Figure 61 - Clustering coefficient histogram for Human H3N2 Dataset

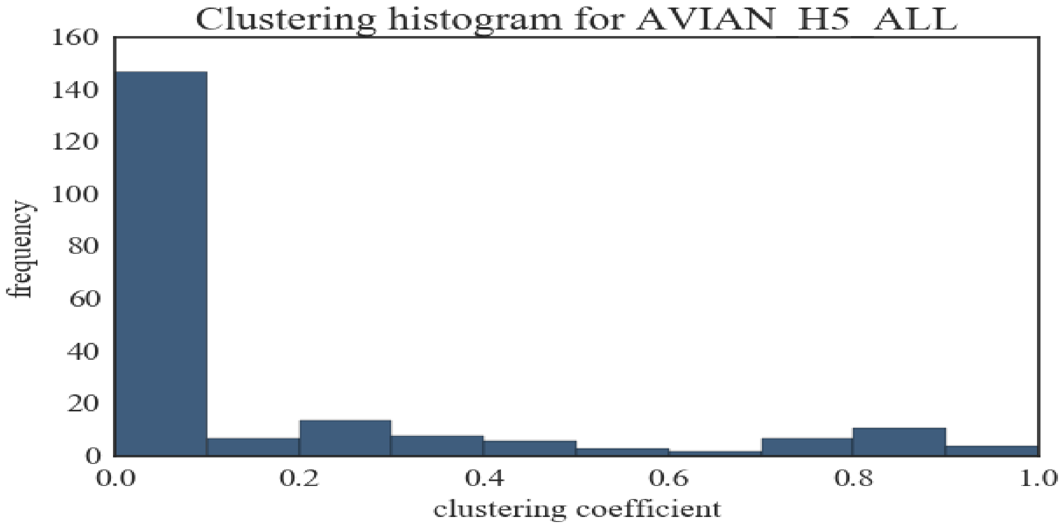


Figure 62 - Clustering coefficient histogram for Avian H5

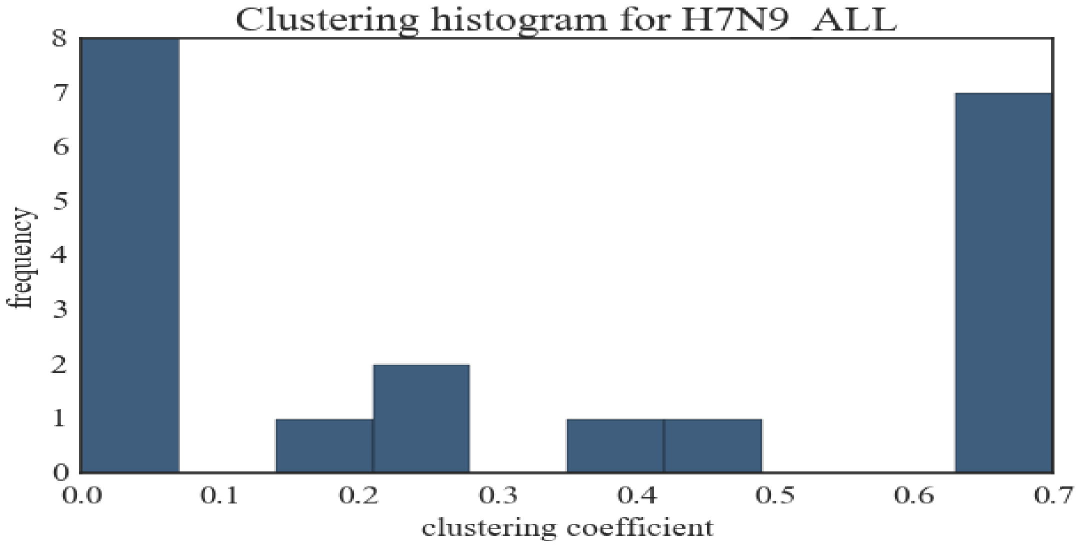


Figure 63 - Clustering coefficient histogram for H7N9 dataset

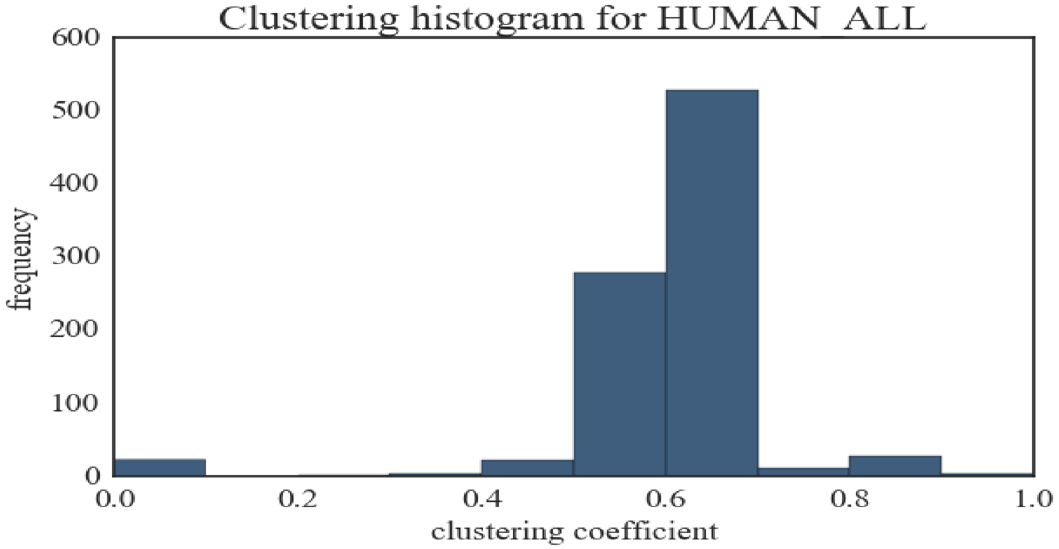


Figure 64 - Clustering coefficient distribution for Human-All dataset

Top 25 Nodes

Nodes with highest degrees in a dataset have a special significance since they tend to act as 'hubs' in the network. We list the top 25 nodes with highest degree distributions for each the seven datasets in Table 8, Table 9, Table 10, Table 11, Table 12, Table 13 and Table 14. We also identify the top 10 edges (covarying residues) for these top nodes. Several interesting observations can be derived from these results.

1. It can be generally concluded that majority of these top nodes are not from virally active surface proteins (HA and NA).
2. The top nodes in Human H1N1 influenza have significantly higher number of connections compared to other datasets. There is also a very small variance between the degrees of top 25 nodes in Human H1N1 dataset. The top node (NS2_63) has a degree of 416 while the 25th node (M2_83) has a degree of 415.
3. In the Avian H5 dataset, HA_217 acts as a hub node with a relatively high degree compared to other top nodes.
4. In the Swine H1N1 dataset, NS2_40 and NS1_197 have relatively higher degrees compared to other top nodes.

Table 8 – Top 25 residues for Swine H1N1 Dataset

RESIDUE	DEGREE	COVARYING_RESIDUES
NS2_40 (Ile)	73	['NS1197 (Asn)', 'NP293 (Ala)', 'PB2705 (Asp)', 'NP27 (Asp)', 'NP145 (Ile)', 'NP465 (Leu)', 'NP393 (Arg)', 'PA748 (Thr)', 'NP108 (Arg)', 'NP380 (Val)']
NS1_197 (Asn)	70	['NS240 (Ile)', 'NP293 (Ala)', 'PB2705 (Asp)', 'NP27 (Asp)', 'NP145 (Ile)', 'PA748 (Thr)', 'NP393 (Arg)', 'NP465 (Leu)', 'NP108 (Arg)', 'NP380 (Val)']
PB2_705 (Asp)	58	['NP293 (Ala)', 'NP393 (Arg)', 'NP108 (Arg)', 'NP145 (Ile)', 'NP27 (Asp)', 'PA748 (Thr)', 'NP465 (Leu)', 'NP380 (Val)', 'NS240 (Ile)', 'PA216 (Asp)']
NP_293 (Ala)	48	['PB2705 (Asp)', 'PA748 (Thr)', 'NS240 (Ile)', 'PA216 (Asp)', 'NS1197 (Asn)', 'PB2483 (Met)', 'PA184 (Ser)', 'PA553 (Ala)', 'PA399 (Glu)', 'PB1592 (Asp)']
NP_298 (His)	44	['PB2705 (Asp)', 'PA748 (Thr)', 'NS240 (Ile)', 'HA256 (Pro)', 'HA399 (Thr)', 'NS1197 (Asn)', 'HA473 (Arg)', 'HA463 (Ser)', 'HA98 (Val)', 'HA10 (Leu)']
NP_27 (Asp)	44	['PB2705 (Asp)', 'PA748 (Thr)', 'NS240 (Ile)', 'NS1197 (Asn)', 'PA553 (Ala)', 'PA216 (Asp)', 'PA399 (Glu)', 'PA184 (Ser)', 'PB2483 (Met)', 'M255 (Phe)']
NP_145 (Ile)	43	['PB2705 (Asp)', 'PA748 (Thr)', 'NS240 (Ile)', 'NS1197 (Asn)', 'PA553 (Ala)', 'PA399 (Glu)', 'PA216 (Asp)', 'PB2483 (Met)', 'PA184 (Ser)', 'M255 (Phe)']
HA_79 (Ile)	43	['NA114 (Val)', 'NA311 (Glu)', 'NA463 (Glu)', 'NA23 (Leu)', 'NA415 (Leu)', 'NA3 (Pro)', 'NA430 (Arg)', 'NA298 (Gly)', 'NA32 (Ile)', 'NA328 (Pro)']
M2_19 (Cys)	42	['M1248 (Met)', 'NS240 (Ile)', 'NP54 (Lys)', 'NS1197 (Asn)', 'NP332 (Ala)', 'NP107 (Arg)', 'PA553 (Ala)', 'NP293 (Ala)', 'PB2705 (Asp)', 'NS252 (Met)']
NP_465 (Leu)	41	['PB2705 (Asp)', 'PA748 (Thr)', 'NS240 (Ile)', 'NS1197 (Asn)', 'PA553 (Ala)', 'PA216 (Asp)', 'PA399 (Glu)', 'PB2483 (Met)', 'PA184 (Ser)', 'M255 (Phe)']
M1_248 (Met)	41	['M219 (Cys)', 'NS240 (Ile)', 'NP54 (Lys)', 'NP332 (Ala)', 'NS1197 (Asn)', 'NP107 (Arg)', 'PA553 (Ala)', 'NP293 (Ala)', 'PB2705 (Asp)', 'PA184 (Ser)']
PB2_483 (Met)	40	['NP293 (Ala)', 'NP108 (Arg)', 'NP393 (Arg)', 'NS240 (Ile)', 'NP145 (Ile)', 'NS1197 (Asn)', 'PA748 (Thr)', 'NP27 (Asp)', 'NP465 (Leu)', 'NP54 (Lys)']
PA_748 (Thr)	37	['NP293 (Ala)', 'PB2705 (Asp)', 'NP145 (Ile)', 'NP393 (Arg)', 'NP27 (Asp)', 'NP108 (Arg)', 'NP465 (Leu)', 'NP380 (Val)', 'NS1197 (Asn)', 'NS240 (Ile)']
NP_386 (Asn)	36	['PB1779 (Ser)', 'PB2271 (Ala)', 'PB2147 (Thr)', 'PB1649 (Asp)', 'PA400 (Pro)', 'PB2591 (Arg)', 'PB1350 (Met)', 'PB2590 (Ser)', 'PB265 (Asp)', 'PB1595 (Gln)']
HA_496 (Asn)	35	['NA395 (Gly)', 'NA385 (Ser)', 'NA84 (Lys)', 'NA311 (Glu)', 'NA435 (Asn)', 'NA331 (Gly)', 'NA287 (Glu)', 'NA285 (Ser)', 'NA430 (Arg)', 'NA166 (Val)']

NP_380 (Val)	35	['PB2705 (Asp)', 'PA748 (Thr)', 'NS240 (Ile)', 'NS1197 (Asn)', 'PA216 (Asp)', 'PA553 (Ala)', 'PA399 (Glu)', 'PA184 (Ser)', 'PB2483 (Met)', 'M255 (Phe)']
PB2_243 (Met)	34	['PB1484 (Val)', 'PA717 (Ala)', 'PB1777 (Glu)', 'PB1394 (Glu)', 'PB1587 (Leu)', 'PA332 (Pro)', 'PB1140 (Tyr)', 'PB191 (Ser)', 'PB1411 (Thr)', 'PA251 (Lys)']
PA_400 (Pro)	34	['PB1779 (Ser)', 'PB2147 (Thr)', 'PB2271 (Ala)', 'PB1649 (Asp)', 'NP386 (Asn)', 'PB2591 (Arg)', 'PB1350 (Met)', 'PB1347 (Ile)', 'PB2590 (Ser)', 'PB265 (Asp)']
PB2_265 (Asn)	33	['PB1484 (Val)', 'PA717 (Ala)', 'PB1777 (Glu)', 'PB1394 (Glu)', 'PB1587 (Leu)', 'PB1140 (Tyr)', 'PB191 (Ser)', 'PB1411 (Thr)', 'PA251 (Lys)', 'PA332 (Pro)']
PB2_467 (Met)	33	['PB1484 (Val)', 'PA717 (Ala)', 'PB1777 (Glu)', 'PB1394 (Glu)', 'PB1587 (Leu)', 'PA332 (Pro)', 'PB1140 (Tyr)', 'PB191 (Ser)', 'PB1411 (Thr)', 'PA251 (Lys)']
PA_256 (Arg)	32	['PB1779 (Ser)', 'PB2147 (Thr)', 'PB2271 (Ala)', 'PB1649 (Asp)', 'NP386 (Asn)', 'PB1350 (Met)', 'PB2591 (Arg)', 'PB2590 (Ser)', 'PB1595 (Gln)', 'PB265 (Asp)']
PA_553 (Ala)	32	['NP107 (Arg)', 'NP332 (Ala)', 'NP54 (Lys)', 'NP293 (Ala)', 'PB2705 (Asp)', 'NP145 (Ile)', 'NP27 (Asp)', 'NP393 (Arg)', 'NP108 (Arg)', 'NP465 (Leu)']
PB1_484 (Val)	32	['PB2627 (Glu)', 'PB2467 (Met)', 'PB2265 (Asn)', 'PB2243 (Met)', 'PA717 (Ala)', 'PB2475 (Leu)', 'PB2199 (Ala)', 'PA542 (Val)', 'PB2238 (Thr)', 'PA332 (Pro)']
NP_393 (Arg)	32	['PB2705 (Asp)', 'PA748 (Thr)', 'NS240 (Ile)', 'NS1197 (Asn)', 'PA216 (Asp)', 'PB2483 (Met)', 'PA184 (Ser)', 'PA553 (Ala)', 'PA399 (Glu)', 'PB1592 (Asp)']
PB2_147 (Thr)	31	['PB1779 (Ser)', 'PB1649 (Asp)', 'PB1350 (Met)', 'PA400 (Pro)', 'NP386 (Asn)', 'PB1595 (Gln)', 'PB1347 (Ile)', 'PA407 (Val)', 'PB1163 (Ser)', 'PA256 (Arg)']

Table 9 - Top 25 nodes for Human H1N1 Dataset

RESIDUE	DEGREE	COVARYING_RESIDUES
NS2_63 (Glu)	419	['PA336 (Met)', 'PB2684 (Ser)', 'PB1584 (Gln)', 'NS1119 (Leu)', 'NS191 (Ser)', 'NP456 (Leu)', 'NP289 (His)', 'NP21 (Asp)', 'HA308 (Thr)', 'HA256 (Lys)']
NS2_107 (Leu)	418	['PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)', 'NP257 (Ile)', 'M286 (Val)', 'M236 (Leu)', 'HA206 (Gln)', 'PA421 (Ser)']
NS2_86 (Arg)	418	['PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)', 'NP257 (Ile)', 'M286 (Val)', 'M236 (Leu)', 'HA206 (Gln)', 'PA421 (Ser)']
NS2_60 (Ser)	418	['PA336 (Met)', 'PB2684 (Ser)', 'PB1584 (Gln)', 'NS1119 (Leu)', 'NS191 (Ser)', 'NP456 (Leu)', 'NP289 (His)', 'NP21 (Asp)', 'HA308 (Thr)', 'HA256 (Lys)']
NS2_40 (Ile)	418	['NS1197 (Asn)', 'PA336 (Met)', 'PB2684 (Ser)', 'PB1584 (Gln)', 'NS1119 (Leu)', 'NS191 (Ser)', 'NP456 (Leu)', 'NP289 (His)', 'NP21 (Asp)', 'HA308 (Thr)']
NS2_57 (Tyr)	417	['PA65 (Ser)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)', 'NS186 (Thr)', 'NP421 (Glu)', 'HA328 (Lys)', 'HA210 (Gln)', 'PB2399 (Ile)', 'PB1691 (Lys)']

M1_121 (Thr)	416	['PB2674 (Ala)', 'PB2645 (Leu)', 'M277 (Gln)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)', 'PA65 (Ser)']
M1_137 (Thr)	416	['PB2674 (Ala)', 'PB2645 (Leu)', 'M277 (Gln)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)', 'PA65 (Ser)']
M1_115 (Val)	416	['NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'NA250 (Gln)', 'NA214 (Asp)', 'NA166 (Val)', 'NA21 (Asn)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)']
M1_209 (Thr)	416	['PB2674 (Ala)', 'PB2645 (Leu)', 'M277 (Gln)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)', 'PA65 (Ser)']
M1_116 (Ser)	416	['PB2674 (Ala)', 'PB2645 (Leu)', 'M277 (Gln)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB1576 (Leu)', 'M278 (Gln)', 'NA250 (Gln)', 'NA214 (Asp)']
M1_231 (Asp)	416	['NA101 (Ser)', 'NA14 (Cys)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)']
M1_160 (Arg)	416	['NA101 (Ser)', 'NA14 (Cys)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)']
M1_101 (Lys)	416	['NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'NA250 (Gln)', 'NA214 (Asp)', 'NA166 (Val)', 'NA21 (Asn)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)']
M1_147 (Val)	416	['NA101 (Ser)', 'NA14 (Cys)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)']
NS2_26 (Glu)	416	['PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)', 'NP257 (Ile)', 'M286 (Val)', 'M236 (Leu)', 'HA206 (Gln)', 'PA404 (Ala)']
M1_142 (Ala)	416	['PB2674 (Ala)', 'PB2645 (Leu)', 'M277 (Gln)', 'NA351 (Phe)', 'NA311 (Glu)', 'NA70 (Ser)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)', 'PB1576 (Leu)']
M1_218 (Thr)	416	['M257 (Tyr)', 'NA101 (Ser)', 'NA14 (Cys)', 'PB2674 (Ala)', 'PB2645 (Leu)', 'PB2399 (Ile)', 'PB2199 (Ala)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)']
NS2_6 (Met)	415	['NS16 (Met)', 'NP190 (Ala)', 'HA385 (Leu)', 'PA336 (Met)', 'PB2684 (Ser)', 'PB1584 (Gln)', 'NS1119 (Leu)', 'NS191 (Ser)', 'NP456 (Leu)', 'NP289 (His)']
M2_57 (Tyr)	415	['PB2674 (Ala)', 'PB2645 (Leu)', 'PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)', 'NP257 (Ile)', 'HA206 (Gln)', 'PB1576 (Leu)']
M2_78 (Gln)	415	['PB2674 (Ala)', 'PB2645 (Leu)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)', 'PA65 (Ser)', 'NS186 (Thr)', 'NP421 (Glu)', 'HA328 (Lys)', 'HA210 (Gln)']
M2_36 (Leu)	415	['PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)', 'NP257 (Ile)', 'HA206 (Gln)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2490 (Ser)']
M2_77 (Gln)	415	['PB2674 (Ala)', 'PB2645 (Leu)', 'PB1576 (Leu)', 'M1209 (Thr)', 'M1137 (Thr)', 'M1121 (Thr)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2114 (Val)', 'PA65 (Ser)']
NS2_115 (Ala)	415	['NS16 (Met)', 'NP190 (Ala)', 'HA385 (Leu)', 'PA336 (Met)', 'PB2684 (Ser)', 'PB1584 (Gln)', 'NS1119 (Leu)', 'NS191 (Ser)', 'NP456 (Leu)', 'NP289 (His)']
M2_86 (Val)	415	['PB2399 (Ile)', 'PB1691 (Lys)', 'NP422 (Arg)', 'NP334 (His)', 'NP293 (Arg)', 'NP257 (Ile)', 'HA206 (Gln)', 'PB2627 (Glu)', 'PB2491 (Thr)', 'PB2490 (Ser)']

Table 10 - Top 25 nodes for Human H3N2 Dataset

RESIDUE	DEGREE	COVARYING_RESIDUES
PB1_709 (Ile)	50	['NP406 (Thr)', 'NA50 (Phe)', 'NA31 (Phe)', 'NA26 (Ser)', 'NA394 (Asn)', 'PB2590 (Ser)', 'NP136 (Ile)', 'HA246 (Pro)', 'NA38 (Ile)', 'NA316 (Ile)']
PB2_590 (Ser)	50	['NP406 (Thr)', 'PB1709 (Ile)', 'NA394 (Asn)', 'NA50 (Phe)', 'NA31 (Phe)', 'NA26 (Ser)', 'NP136 (Ile)', 'HA246 (Pro)', 'NA38 (Ile)', 'NA316 (Ile)']

PA_420 (Asp)	49	['PB1113 (Val)', 'M231 (Ser)', 'PB2249 (Glu)', 'PB2451 (Ile)', 'NA396 (Asn)', 'HA69 (Glu)', 'NA158 (His)', 'NP406 (Thr)', 'NA394 (Asn)', 'NA50 (Phe)']
PA_101 (Glu)	47	['PB1113 (Val)', 'M231 (Ser)', 'PB2249 (Glu)', 'PB2451 (Ile)', 'HA69 (Glu)', 'NA396 (Asn)', 'NA158 (His)', 'HA246 (Pro)', 'NP406 (Thr)', 'PB1709 (Ile)']
M2_31 (Ser)	45	['PB1113 (Val)', 'PA101 (Glu)', 'PA420 (Asp)', 'HA69 (Glu)', 'PA475 (Tyr)', 'NA396 (Asn)', 'NA381 (Ser)', 'NA158 (His)', 'PA256 (Lys)', 'HA159 (Lys)']
NP_406 (Thr)	44	['PB1709 (Ile)', 'NA31 (Phe)', 'NA26 (Ser)', 'NA50 (Phe)', 'NA394 (Asn)', 'PB2590 (Ser)', 'HA246 (Pro)', 'NA38 (Ile)', 'NA316 (Ile)', 'M251 (Val)']
M2_51 (Val)	43	['NP406 (Thr)', 'NA26 (Ser)', 'PB1709 (Ile)', 'NA31 (Phe)', 'NA50 (Phe)', 'NA394 (Asn)', 'PB2590 (Ser)', 'NP136 (Ile)', 'NA316 (Ile)', 'HA246 (Pro)']
NP_136 (Ile)	43	['PB1709 (Ile)', 'NA50 (Phe)', 'NA31 (Phe)', 'PB2590 (Ser)', 'NA394 (Asn)', 'NA26 (Ser)', 'HA246 (Pro)', 'NA38 (Ile)', 'NA316 (Ile)', 'M251 (Val)']
PB1_113 (Val)	42	['M231 (Ser)', 'PA101 (Glu)', 'PA420 (Asp)', 'PA475 (Tyr)', 'NA396 (Asn)', 'HA69 (Glu)', 'PA256 (Lys)', 'NA158 (His)', 'HA159 (Lys)', 'NP131 (Ala)']
NA_50 (Phe)	41	['NP406 (Thr)', 'PB1709 (Ile)', 'PB2590 (Ser)', 'NP136 (Ile)', 'HA246 (Pro)', 'PB1619 (Asn)', 'M251 (Val)', 'PB2340 (Lys)', 'NS182 (Val)', 'NP77 (Lys)']
NS1_82 (Val)	41	['NA26 (Ser)', 'NP406 (Thr)', 'NA31 (Phe)', 'PB1709 (Ile)', 'PB1586 (Arg)', 'NA50 (Phe)', 'NA394 (Asn)', 'NP136 (Ile)', 'PB2590 (Ser)', 'NA316 (Ile)']
NA_394 (Asn)	40	['NP406 (Thr)', 'PB1709 (Ile)', 'PB2590 (Ser)', 'NP136 (Ile)', 'HA246 (Pro)', 'PB1619 (Asn)', 'M251 (Val)', 'PB2340 (Lys)', 'NS182 (Val)', 'HA221 (Ile)']
NP_312 (Val)	40	['HA246 (Pro)', 'PB1709 (Ile)', 'NA26 (Ser)', 'NA50 (Phe)', 'NA31 (Phe)', 'NA394 (Asn)', 'PB2590 (Ser)', 'NA316 (Ile)', 'NA38 (Ile)', 'NA224 (Val)']
NA_31 (Phe)	40	['NP406 (Thr)', 'PB1709 (Ile)', 'PB2590 (Ser)', 'NP136 (Ile)', 'HA246 (Pro)', 'PB1619 (Asn)', 'M251 (Val)', 'PB2340 (Lys)', 'NS182 (Val)', 'NP77 (Lys)']
NA_26 (Ser)	40	['NP406 (Thr)', 'PB1709 (Ile)', 'PB2590 (Ser)', 'PB1619 (Asn)', 'NP136 (Ile)', 'HA246 (Pro)', 'M251 (Val)', 'PB2340 (Lys)', 'NS182 (Val)', 'NP77 (Lys)']
PB2_340 (Lys)	40	['NA26 (Ser)', 'NP406 (Thr)', 'PB1709 (Ile)', 'NA31 (Phe)', 'NA394 (Asn)', 'NA50 (Phe)', 'NA316 (Ile)', 'NP136 (Ile)', 'PB1619 (Asn)', 'NA38 (Ile)']
NP_52 (Tyr)	39	['HA246 (Pro)', 'PB1709 (Ile)', 'PB2249 (Glu)', 'NA26 (Ser)', 'PB2451 (Ile)', 'NA31 (Phe)', 'NA50 (Phe)', 'NA394 (Asn)', 'PB2590 (Ser)', 'HA245 (Ile)']
HA_69 (Glu)	39	['NA381 (Ser)', 'NA396 (Asn)', 'NA158 (His)', 'NP131 (Ala)', 'PB2249 (Glu)', 'M231 (Ser)', 'PB2451 (Ile)', 'PB1113 (Val)', 'NA319 (Tyr)', 'NS1221 (Lys)']
PB1_619 (Asn)	38	['NA26 (Ser)', 'NP406 (Thr)', 'NA31 (Phe)', 'NA50 (Phe)', 'NA394 (Asn)', 'NA316 (Ile)', 'PB2590 (Ser)', 'NP136 (Ile)', 'HA246 (Pro)', 'NA38 (Ile)']
HA_246 (Pro)	38	['PB1709 (Ile)', 'NP406 (Thr)', 'NA50 (Phe)', 'NA26 (Ser)', 'NA31 (Phe)', 'NA394 (Asn)', 'PB2590 (Ser)', 'NP136 (Ile)', 'NA38 (Ile)', 'NA316 (Ile)']
PB2_249 (Glu)	38	['NA396 (Asn)', 'NP131 (Ala)', 'HA159 (Lys)', 'NA158 (His)', 'HA69 (Glu)', 'NA381 (Ser)', 'NA319 (Tyr)', 'PA420 (Asp)', 'NA202 (Val)', 'PA101 (Glu)']
NA_38 (Ile)	37	['NP406 (Thr)', 'PB1709 (Ile)', 'PB2590 (Ser)', 'NP136 (Ile)', 'HA246 (Pro)', 'PB1619 (Asn)', 'M251 (Val)', 'PB2340 (Lys)', 'NS182 (Val)', 'NP77 (Lys)']
NP_280 (Val)	36	['HA246 (Pro)', 'PB1709 (Ile)', 'PB2590 (Ser)', 'NA26 (Ser)', 'NA50 (Phe)', 'NA31 (Phe)', 'NA394 (Asn)', 'NA316 (Ile)', 'NA224 (Val)', 'NA38 (Ile)']
NA_316 (Ile)	36	['NP406 (Thr)', 'PB1709 (Ile)', 'PB2590 (Ser)', 'PB1619 (Asn)', 'NP136 (Ile)', 'HA246 (Pro)', 'M251 (Val)', 'PB2340 (Lys)', 'NS182 (Val)', 'NP77 (Lys)']
PB2_451 (Ile)	36	['NA396 (Asn)', 'HA159 (Lys)', 'NP131 (Ala)', 'NA158 (His)', 'HA69 (Glu)', 'NA381 (Ser)', 'NA319 (Tyr)', 'NS1221 (Lys)', 'PA420 (Asp)', 'NA202 (Val)']

Table 11 - Top 25 residues for Swine H3N2 Dataset

RESIDUE	DEGREE	COVARYING_RESIDUES
M2_28 (Ile)	13	['M195 (Arg)', 'M1214 (His)', 'M1121 (Thr)', 'M1116 (Ser)', 'M1101 (Lys)', 'M1166 (Ala)', 'M115 (Ile)', 'M1142 (Ala)', 'M1209 (Thr)', 'M130 (Ser)']
M2_20 (Ser)	13	['M1214 (His)', 'M1121 (Thr)', 'M1116 (Ser)', 'M1166 (Ala)', 'M1101 (Lys)', 'M195 (Arg)', 'M1142 (Ala)', 'M1209 (Thr)', 'M115 (Ile)', 'M1207 (Asn)']
M2_31 (Asn)	13	['M1116 (Ser)', 'M1214 (His)', 'M1121 (Thr)', 'M1166 (Ala)', 'M1101 (Lys)', 'M195 (Arg)', 'M1209 (Thr)', 'M1142 (Ala)', 'M115 (Ile)', 'M1207 (Asn)']
M2_95 (Glu)	13	['M1214 (His)', 'M1121 (Thr)', 'M1116 (Ser)', 'M195 (Arg)', 'M1166 (Ala)', 'M1101 (Lys)', 'M1209 (Thr)', 'M1142 (Ala)', 'M115 (Ile)', 'M1181 (Leu)']
M2_77 (Gln)	13	['M1116 (Ser)', 'M1214 (His)', 'M1121 (Thr)', 'M1101 (Lys)', 'M1166 (Ala)', 'M195 (Arg)', 'M1142 (Ala)', 'M1209 (Thr)', 'M115 (Ile)', 'M1207 (Asn)']
M2_60 (Lys)	13	['M1181 (Leu)', 'M115 (Ile)', 'M1139 (Thr)', 'M1121 (Thr)', 'M1214 (His)', 'M1116 (Ser)', 'M1101 (Lys)', 'M1166 (Ala)', 'M195 (Arg)', 'M1209 (Thr)']
M2_18 (Arg)	13	['M1181 (Leu)', 'M1214 (His)', 'M1121 (Thr)', 'M195 (Arg)', 'M1166 (Ala)', 'M1116 (Ser)', 'M1207 (Asn)', 'M130 (Ser)', 'M1209 (Thr)', 'M1101 (Lys)']
M2_79 (Glu)	11	['M1214 (His)', 'M1121 (Thr)', 'M1116 (Ser)', 'M1166 (Ala)', 'M195 (Arg)', 'M1101 (Lys)', 'M1209 (Thr)', 'M1142 (Ala)', 'M115 (Ile)', 'M1207 (Asn)']
M2_43 (Thr)	11	['M130 (Ser)', 'M1207 (Asn)', 'M1142 (Ala)', 'M1209 (Thr)', 'M1166 (Ala)', 'M1116 (Ser)', 'M1214 (His)', 'M1121 (Thr)', 'M1101 (Lys)', 'M195 (Arg)']
M1_207 (Asn)	10	['M243 (Thr)', 'M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M214 (Glu)', 'M260 (Lys)', 'M218 (Arg)', 'M279 (Glu)']
M1_30 (Ser)	10	['M243 (Thr)', 'M214 (Glu)', 'M228 (Ile)', 'M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M218 (Arg)', 'M260 (Lys)', 'M213 (Ser)']
M1_142 (Ala)	10	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M243 (Thr)', 'M260 (Lys)', 'M279 (Glu)', 'M218 (Arg)', 'M214 (Glu)']
M1_209 (Thr)	10	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M260 (Lys)', 'M243 (Thr)', 'M279 (Glu)', 'M218 (Arg)', 'M214 (Glu)']
M1_116 (Ser)	9	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M260 (Lys)', 'M279 (Glu)', 'M243 (Thr)', 'M218 (Arg)', 'M214 (Glu)']
M1_95 (Arg)	9	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M260 (Lys)', 'M243 (Thr)', 'M279 (Glu)', 'M218 (Arg)', 'M214 (Glu)']
M1_121 (Thr)	9	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M260 (Lys)', 'M279 (Glu)', 'M243 (Thr)', 'M218 (Arg)', 'M214 (Glu)']
M1_214 (His)	9	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M260 (Lys)', 'M279 (Glu)', 'M243 (Thr)', 'M218 (Arg)', 'M214 (Glu)']
M1_166 (Ala)	9	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M260 (Lys)', 'M243 (Thr)', 'M279 (Glu)', 'M218 (Arg)', 'M214 (Glu)']
M1_101 (Lys)	9	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M295 (Glu)', 'M228 (Ile)', 'M260 (Lys)', 'M243 (Thr)', 'M279 (Glu)', 'M218 (Arg)', 'M214 (Glu)']
M1_181 (Leu)	9	['M260 (Lys)', 'M295 (Glu)', 'M220 (Ser)', 'M277 (Gln)', 'M231 (Asn)', 'M228 (Ile)', 'M218 (Arg)', 'M214 (Glu)', 'M279 (Glu)', 'M243 (Thr)']
M1_15 (Ile)	9	['M277 (Gln)', 'M220 (Ser)', 'M231 (Asn)', 'M260 (Lys)', 'M295 (Glu)', 'M228 (Ile)', 'M243 (Thr)', 'M279 (Glu)', 'M218 (Arg)', 'M214 (Glu)']
M1_139 (Thr)	7	['M260 (Lys)', 'M295 (Glu)', 'M220 (Ser)', 'M277 (Gln)', 'M231 (Asn)', 'M228 (Ile)', 'M218 (Arg)', 'M214 (Glu)', 'M279 (Glu)', 'M243 (Thr)']
M2_14 (Glu)	5	['M130 (Ser)', 'M1207 (Asn)', 'M1209 (Thr)', 'M1142 (Ala)', 'M1181 (Leu)', 'M1166 (Ala)', 'M1116 (Ser)', 'M1214 (His)', 'M1121 (Thr)', 'M195 (Arg)']
NS2_37 (Ala)	4	['NS1194 (Gly)', 'NS1206 (Arg)', 'NS191 (Ala)', 'NS1171 (Asn)', 'NS1129 (Ile)', 'HA246 (Ile)', 'NS1209 (Asp)', 'HA243 (Val)', 'NA395 (Thr)', 'NA16 (Ile)']

NS1_70 (Lys)	2	['HA180 (Ala)', 'HA507 (Val)', 'NS214 (Met)', 'HA220 (Thr)', 'HA316 (Arg)', 'PB261 (Lys)', 'HA124 (Ser)', 'PB2225 (Gly)', 'HA134 (Thr)', 'NP223 (Ile)']

Table 12 - Top 25 residues for Avian H5 Dataset

RESIDUE	DEGREE	COVARYING_RESIDUES
HA_217 (Ile)	89	['NA34 (Met)', 'NA26 (Met)', 'NA93 (Ser)', 'NA134 (Val)', 'NA430 (Ser)', 'NA256 (Asn)', 'NA461 (Leu)', 'NA452 (Val)', 'NA437 (Asp)', 'NA223 (Thr)']
PB1_384 (Ser)	66	['NA34 (Met)', 'NA343 (Ser)', 'NA256 (Asn)', 'NA461 (Leu)', 'NA334 (His)', 'NA25 (Cys)', 'NA452 (Val)', 'NA437 (Asp)', 'NA430 (Ser)', 'NA161 (His)']
PB1_149 (Ile)	43	['NS1219 (Lys)', 'NA204 (Pro)', 'NA203 (Ser)', 'PB272 (Met)', 'NS184 (Met)', 'NA325 (Ile)', 'NA32 (Ser)', 'NP483 (Asn)', 'NA519 (Glu)', 'NP377 (Asn)']
NS1_84 (Met)	35	['NP377 (Asn)', 'NS222 (Ala)', 'NS256 (Phe)', 'NP483 (Asn)', 'PB114 (Val)', 'M127 (Lys)', 'PB1149 (Ile)', 'NA325 (Ile)', 'NS261 (Ile)', 'NA32 (Ser)']
PB1_14 (Val)	31	['NP483 (Asn)', 'NS184 (Met)', 'NP377 (Asn)', 'NS1219 (Lys)', 'NS1214 (Leu)', 'NS222 (Ala)', 'NA325 (Ile)', 'NS256 (Phe)', 'NA204 (Pro)', 'NA203 (Ser)']
NP_377 (Asn)	31	['NS184 (Met)', 'PB114 (Val)', 'NS222 (Ala)', 'NS1219 (Lys)', 'NS1214 (Leu)', 'NS256 (Phe)', 'PB1149 (Ile)', 'NA204 (Pro)', 'NA203 (Ser)', 'NA519 (Glu)']
NS1_219 (Lys)	31	['NS222 (Ala)', 'NS256 (Phe)', 'NS261 (Ile)', 'NP377 (Asn)', 'PB1149 (Ile)', 'PB114 (Val)', 'NP483 (Asn)', 'M127 (Lys)', 'NA325 (Ile)', 'NA32 (Ser)']
NP_483 (Asn)	30	['PB114 (Val)', 'NS184 (Met)', 'NS222 (Ala)', 'PB1149 (Ile)', 'NS1214 (Leu)', 'NS1219 (Lys)', 'NS256 (Phe)', 'NA325 (Ile)', 'NA204 (Pro)', 'NA203 (Ser)']
NS2_22 (Ala)	28	['NS1214 (Leu)', 'NS184 (Met)', 'NS1219 (Lys)', 'NP377 (Asn)', 'M127 (Lys)', 'NP483 (Asn)', 'PB114 (Val)', 'NS148 (Asn)', 'PB1149 (Ile)', 'NA325 (Ile)']
NS1_214 (Leu)	25	['NS256 (Phe)', 'NS222 (Ala)', 'NP377 (Asn)', 'NP483 (Asn)', 'M127 (Lys)', 'NS261 (Ile)', 'PB114 (Val)', 'PB1149 (Ile)', 'NA325 (Ile)', 'NA32 (Ser)']
NS2_56 (Phe)	24	['NS1214 (Leu)', 'NS1219 (Lys)', 'NS184 (Met)', 'NP377 (Asn)', 'NP483 (Asn)', 'M127 (Lys)', 'PB114 (Val)', 'NS148 (Asn)', 'PB1149 (Ile)', 'NA325 (Ile)']
M2_14 (Glu)	22	['M1232 (Asn)', 'M1224 (Asn)', 'HA398 (Val)', 'M1207 (Asn)', 'HA126 (Ile)', 'M1166 (Ala)', 'M1230 (Arg)', 'HA125 (Arg)', 'HA236 (Ser)', 'HA418 (Arg)']
M1_224 (Asn)	19	['HA228 (Leu)', 'HA126 (Ile)', 'HA236 (Ser)', 'HA125 (Arg)', 'HA418 (Arg)', 'HA328 (Asn)', 'M214 (Glu)', 'HA398 (Val)', 'HA558 (Leu)', 'HA144 (Glu)']
NS2_69 (Gln)	19	['NS1159 (Gly)', 'NS114 (Phe)', 'NS123 (Ala)', 'NS142 (Ser)', 'NS1164 (Leu)', 'NS1146 (Leu)', 'NS1171 (Thr)', 'NS194 (Thr)', 'NS128 (Gly)', 'NS125 (Gln)']
PB2_72 (Met)	19	['PB1149 (Ile)', 'PB114 (Val)', 'NS184 (Met)', 'NS1219 (Lys)', 'NS222 (Ala)', 'NA204 (Pro)', 'NA203 (Ser)', 'NP377 (Asn)', 'NP483 (Asn)', 'NA32 (Ser)']
NS1_48 (Asn)	18	['NS256 (Phe)', 'NS222 (Ala)', 'NP377 (Asn)', 'M127 (Lys)', 'NA325 (Ile)', 'NP483 (Asn)', 'NS261 (Ile)', 'PB1149 (Ile)', 'NA204 (Pro)', 'NA203 (Ser)']
M1_101 (Lys)	17	['HA126 (Ile)', 'HA349 (Lys)', 'HA236 (Ser)', 'HA125 (Arg)', 'HA418 (Arg)', 'HA398 (Val)', 'HA228 (Leu)', 'HA558 (Leu)', 'M214 (Glu)', 'HA61 (Asp)']
M1_166 (Ala)	17	['HA126 (Ile)', 'HA349 (Lys)', 'HA125 (Arg)', 'HA236 (Ser)', 'HA418 (Arg)', 'HA398 (Val)', 'HA558 (Leu)', 'HA228 (Leu)', 'M214 (Glu)', 'HA339 (Ser)']
M1_232 (Asn)	16	['M214 (Glu)', 'HA126 (Ile)', 'HA125 (Arg)', 'HA418 (Arg)', 'HA236 (Ser)', 'HA398 (Val)', 'HA558 (Leu)', 'HA228 (Leu)', 'HA349 (Lys)', 'HA144 (Glu)']
M1_15 (Ile)	15	['HA126 (Ile)', 'HA236 (Ser)', 'HA418 (Arg)', 'HA125 (Arg)', 'HA398 (Val)', 'HA228 (Leu)', 'HA558 (Leu)', 'M214 (Glu)', 'HA339 (Ser)', 'HA61 (Asp)']

M1_207 (Asn)	14	['HA61 (Asp)', 'HA126 (Ile)', 'M214 (Glu)', 'HA125 (Arg)', 'HA236 (Ser)', 'HA418 (Arg)', 'HA398 (Val)', 'HA558 (Leu)', 'HA228 (Leu)', 'HA349 (Lys)']
NS2_38 (Ser)	14	['NS1159 (Gly)', 'NS114 (Phe)', 'NS142 (Ser)', 'NS1171 (Thr)', 'NS194 (Thr)', 'NS128 (Gly)', 'NS133 (Leu)', 'NS123 (Ala)', 'NS1164 (Leu)', 'NS1146 (Leu)']
NA_325 (Ile)	14	['PB1149 (Ile)', 'NS184 (Met)', 'NP483 (Asn)', 'PB114 (Val)', 'NS1219 (Lys)', 'NP377 (Asn)', 'NS222 (Ala)', 'NS1214 (Leu)', 'NS256 (Phe)', 'NS148 (Asn)']
NP_34 (Ser)	14	['NS184 (Met)', 'PB114 (Val)', 'NS1219 (Lys)', 'NS1214 (Leu)', 'NS222 (Ala)', 'NS256 (Phe)', 'NA223 (Thr)', 'PB1149 (Ile)', 'M127 (Lys)', 'NA519 (Glu)']
M1_27 (Lys)	14	['NS184 (Met)', 'NS222 (Ala)', 'NS1214 (Leu)', 'NS256 (Phe)', 'NS1219 (Lys)', 'NP377 (Asn)', 'PB114 (Val)', 'NP483 (Asn)', 'PB1149 (Ile)', 'NS148 (Asn)']

Table 13 - Top 25 residues for H7N9 Dataset

RESIDUE	DEGREE	COVARYING_RESIDUES
PB1_171 (Met)	10	['NP375 (Asp)', 'NS1152 (Glu)', 'NA83 (Arg)', 'PB2139 (Val)', 'NS127 (Leu)', 'NP371 (Met)', 'NS1216 (Thr)', 'NS1111 (Ile)', 'NS180 (Ser)', 'NS1212 (Ser)']
PB1_397 (Ile)	10	['NP375 (Asp)', 'NS1152 (Glu)', 'NA83 (Arg)', 'NS1216 (Thr)', 'NS1111 (Ile)', 'NS180 (Ser)', 'NS127 (Leu)', 'NP371 (Met)', 'PB2139 (Val)', 'NS1212 (Ser)']
NP_375 (Asp)	9	['PB1397 (Ile)', 'PB1171 (Met)', 'NA83 (Arg)', 'NS1152 (Glu)', 'NS1111 (Ile)', 'NS1216 (Thr)', 'NS127 (Leu)', 'NS180 (Ser)', 'NS1212 (Ser)', 'PB2139 (Val)']
PB2_139 (Val)	5	['PB1694 (Asn)', 'PB1397 (Ile)', 'PB1171 (Met)', 'HA140 (Thr)', 'NS1152 (Glu)', 'NS227 (Asp)', 'NP375 (Asp)', 'NS127 (Leu)', 'NS1216 (Thr)', 'NS1111 (Ile)']
NS1_152 (Glu)	5	['PB1397 (Ile)', 'PB1171 (Met)', 'NP375 (Asp)', 'NA83 (Arg)', 'PB2139 (Val)', 'NP371 (Met)', 'PB1694 (Asn)', 'HA140 (Thr)', 'PB2676 (Met)', 'NS227 (Asp)']
NA_83 (Arg)	5	['PB1397 (Ile)', 'NP375 (Asp)', 'PB1171 (Met)', 'NS1152 (Glu)', 'NP371 (Met)', 'NS227 (Asp)', 'NS1216 (Thr)', 'NS1111 (Ile)', 'NS180 (Ser)', 'NS127 (Leu)']
NS1_216 (Thr)	3	['PB1397 (Ile)', 'PB1171 (Met)', 'NP375 (Asp)', 'PB2676 (Met)', 'NA83 (Arg)', 'PB2139 (Val)', 'PB1694 (Asn)', 'NP371 (Met)', 'PB1525 (Ile)', 'HA140 (Thr)']
NS1_80 (Ser)	3	['PB1397 (Ile)', 'PB1171 (Met)', 'NP375 (Asp)', 'PB2676 (Met)', 'PB2139 (Val)', 'NA83 (Arg)', 'PB1694 (Asn)', 'NP371 (Met)', 'PB1525 (Ile)', 'HA140 (Thr)']
NS1_27 (Leu)	3	['PB1397 (Ile)', 'PB1171 (Met)', 'NP375 (Asp)', 'PB2139 (Val)', 'PB2676 (Met)', 'PB1694 (Asn)', 'NA83 (Arg)', 'NP371 (Met)', 'PB1525 (Ile)', 'HA140 (Thr)']
NS1_111 (Ile)	3	['PB1397 (Ile)', 'NP375 (Asp)', 'PB1171 (Met)', 'PB2139 (Val)', 'NA83 (Arg)', 'PB2676 (Met)', 'PB1694 (Asn)', 'NP371 (Met)', 'PB1525 (Ile)', 'HA140 (Thr)']
NP_371 (Met)	3	['PB1397 (Ile)', 'PB1171 (Met)', 'NA83 (Arg)', 'NS1152 (Glu)', 'PB1525 (Ile)', 'NS1111 (Ile)', 'PB2139 (Val)', 'NS180 (Ser)', 'NS127 (Leu)', 'NS1216 (Thr)']
NS1_212 (Ser)	3	['PB1397 (Ile)', 'NP375 (Asp)', 'PB1171 (Met)', 'PB2676 (Met)', 'PB2139 (Val)', 'NA83 (Arg)', 'PB1694 (Asn)', 'NP371 (Met)', 'PB1525 (Ile)', 'HA140 (Thr)']
NA_327 (Asn)	3	['PB2647 (Ile)', 'PB2535 (Met)', 'PB2511 (Val)', 'M210 (Pro)', 'M224 (Glu)', 'NP371 (Met)', 'PB1397 (Ile)', 'NP375 (Asp)', 'PB1171 (Met)', 'PB2139 (Val)']
PB1_694 (Asn)	2	['PB2139 (Val)', 'PB2676 (Met)', 'NS227 (Asp)', 'HA140 (Thr)', 'NS127 (Leu)', 'NS1216 (Thr)', 'NS1111 (Ile)', 'NS180 (Ser)', 'NP375 (Asp)', 'NS1152 (Glu)']
PB2_535 (Met)	1	['NA327 (Asn)', 'NA241 (Val)', 'M210 (Pro)', 'NA465 (Lys)', 'M224 (Glu)', 'NA238 (Val)', 'PB1397 (Ile)', 'NP371 (Met)', 'PB1171 (Met)', 'NP375 (Asp)']
PB2_676 (Met)	1	['PB1694 (Asn)', 'NS1216 (Thr)', 'NS127 (Leu)', 'NS180 (Ser)', 'NS1212 (Ser)', 'NS1111 (Ile)', 'HA140 (Thr)', 'PB1397 (Ile)', 'PB1171 (Met)', 'NS227 (Asp)']
HA_140 (Thr)	1	['PB2139 (Val)', 'PB1694 (Asn)', 'PB2676 (Met)', 'PB1397 (Ile)', 'PB1171 (Met)', 'NS1152 (Glu)', 'NS127 (Leu)', 'NS1216 (Thr)', 'NS1111 (Ile)', 'NS180 (Ser)']

PB2_511 (Val)	1	['NA327 (Asn)', 'NA241 (Val)', 'M210 (Pro)', 'NA465 (Lys)', 'M224 (Glu)', 'NA238 (Val)', 'PB1397 (Ile)', 'NP371 (Met)', 'PB1171 (Met)', 'NP375 (Asp)']
PB2_647 (Ile)	1	['NA327 (Asn)', 'NA241 (Val)', 'M210 (Pro)', 'NA465 (Lys)', 'M224 (Glu)', 'NA238 (Val)', 'PB1397 (Ile)', 'NP371 (Met)', 'PB1171 (Met)', 'NP375 (Asp)']

Table 14 - Top 25 residues for Human-All Dataset

RESIDUE	DEGREE	COVARYING_RESIDUES
PB1_327 (Lys)	745	['HA163 (Ser)', 'NA136 (Pro)', 'NA57 (Ile)', 'NP439 (Thr)', 'NA206 (Leu)', 'NA48 (Ser)', 'HA187 (Ser)', 'NS2105 (Thr)', 'NA414 (Ser)', 'NP461 (Glu)']
NS2_73 (Leu)	745	['NP429 (Ser)', 'NP465 (Gln)', 'PB2690 (Gly)', 'NS184 (Thr)', 'NA413 (Glu)', 'PB2126 (Glu)', 'NP349 (Val)', 'NS167 (Lys)', 'NA392 (His)', 'HA270 (Ile)']
NS2_76 (Asn)	742	['HA238 (Ser)', 'M243 (Leu)', 'NS1231 (Lys)', 'PB1517 (Val)', 'PA65 (Leu)', 'NA483 (Asn)', 'PB1386 (Lys)', 'NA436 (Ser)', 'M1142 (Val)', 'HA255 (Glu)']
M2_43 (Leu)	742	['HA238 (Ser)', 'NS1231 (Lys)', 'NS276 (Asn)', 'PB1386 (Lys)', 'M1142 (Val)', 'PB1517 (Val)', 'PA65 (Leu)', 'NS144 (Lys)', 'NA436 (Ser)', 'NA400 (Lys)']
M2_78 (Lys)	742	['PA65 (Leu)', 'M1142 (Val)', 'NS1128 (Val)', 'HA255 (Glu)', 'NS144 (Lys)', 'NA96 (Glu)', 'NP350 (Ser)', 'HA119 (Thr)', 'NA490 (Gly)', 'NA105 (Tyr)']
M1_15 (Val)	741	['HA240 (Lys)', 'NP436 (Thr)', 'PB2461 (His)', 'HA284 (Tyr)', 'NA499 (Asn)', 'NA107 (Asn)', 'NS126 (Gly)', 'NS125 (Gln)', 'NS1210 (Glu)', 'M243 (Leu)']
NS2_14 (Met)	740	['NA490 (Gly)', 'NA102 (Asn)', 'NA105 (Tyr)', 'NA49 (Leu)', 'HA415 (Ser)', 'NA333 (Arg)', 'NA301 (Phe)', 'NA263 (Ser)', 'NA232 (Ala)', 'NA110 (Asn)']
M2_89 (Ser)	739	['PA66 (Asp)', 'NP40 (Gly)', 'NA387 (Asn)', 'NS1123 (Glu)', 'PA396 (Asp)', 'HA151 (Ser)', 'PA398 (Lys)', 'PA65 (Leu)', 'PA142 (Lys)', 'NA258 (Thr)']
M1_239 (Ala)	737	['NS1159 (Val)', 'NP292 (Ala)', 'NP379 (Asn)', 'NS141 (Lys)', 'HA190 (Ala)', 'HA272 (Leu)', 'PA633 (Thr)', 'NS128 (Gly)', 'HA339 (Lys)', 'HA67 (Lys)']
PB1_517 (Val)	737	['HA238 (Ser)', 'M243 (Leu)', 'NS276 (Asn)', 'NA483 (Asn)', 'PA65 (Leu)', 'HA255 (Glu)', 'NA436 (Ser)', 'NS1231 (Lys)', 'M1142 (Val)', 'NA400 (Lys)']
M1_142 (Val)	736	['HA255 (Glu)', 'M278 (Lys)', 'NA96 (Glu)', 'HA91 (Ile)', 'NA105 (Tyr)', 'NA102 (Asn)', 'HA119 (Thr)', 'NA110 (Asn)', 'NA49 (Leu)', 'PA65 (Leu)']
PB1_212 (Leu)	736	['NA283 (Phe)', 'NA263 (Ser)', 'NA255 (Ile)', 'NA237 (Gly)', 'NA235 (Thr)', 'NA232 (Ala)', 'NA98 (Thr)', 'HA195 (Tyr)', 'HA152 (Ser)', 'NA333 (Arg)']
M1_205 (Val)	735	['PB2577 (Thr)', 'NS128 (Gly)', 'PA633 (Thr)', 'NP379 (Asn)', 'HA190 (Ala)', 'NP292 (Ala)', 'HA339 (Lys)', 'NS1159 (Val)', 'NA186 (Lys)', 'NS141 (Lys)']
NS2_105 (Thr)	734	['NA206 (Leu)', 'NA48 (Ser)', 'HA187 (Ser)', 'NA414 (Ser)', 'HA163 (Ser)', 'NA136 (Pro)', 'NP299 (Lys)', 'PA417 (Ser)', 'NP448 (Ala)', 'NP378 (Asp)']
M2_54 (Leu)	732	['PA633 (Thr)', 'NS128 (Gly)', 'NP379 (Asn)', 'HA339 (Lys)', 'HA190 (Ala)', 'NS1159 (Val)', 'NP292 (Ala)', 'M1239 (Ala)', 'HA416 (Val)', 'M1205 (Val)']
NS2_123 (Phe)	731	['PA417 (Ser)', 'NP448 (Ala)', 'NP428 (Lys)', 'NP378 (Asp)', 'NP299 (Lys)', 'PA434 (Ser)', 'PA268 (Ile)', 'PB250 (Ser)', 'PA225 (Cys)', 'PB2298 (Thr)']
NS2_6 (Val)	730	['NS16 (Val)', 'PA340 (Leu)', 'NA447 (Glu)', 'NP59 (Glu)', 'PA275 (Pro)', 'PA186 (Gly)', 'NA313 (Leu)', 'NP322 (Ile)', 'PA85 (Thr)', 'PB2231 (Ser)']
PB1_298 (Ile)	729	['NA439 (Lys)', 'NA49 (Leu)', 'NA490 (Gly)', 'NA102 (Asn)', 'HA91 (Ile)', 'NA105 (Tyr)', 'HA415 (Ser)', 'HA119 (Thr)', 'NP350 (Ser)', 'NA111 (Thr)']
PB1_339 (Ile)	729	['NP295 (Tyr)', 'NP196 (Val)', 'PA341 (Ser)', 'NP319 (Tyr)', 'NP67 (Leu)', 'M1207 (Ser)', 'NP289 (Pro)', 'PB2653 (Met)', 'NP27 (Asn)', 'PB2205 (Ser)']
M1_167 (Thr)	729	['NS273 (Leu)', 'NS184 (Thr)', 'NP429 (Ser)', 'NP465 (Gln)', 'PB2690 (Gly)', 'PB2126 (Glu)', 'NS167 (Lys)', 'NA413 (Glu)', 'NP349 (Val)', 'NA392 (His)']

NS2_99 (Val)	728	['NP59 (Glu)', 'NA447 (Glu)', 'PA275 (Pro)', 'PA186 (Gly)', 'NA313 (Leu)', 'HA280 (Ile)', 'NP322 (Ile)', 'PA85 (Thr)', 'M1121 (Ala)', 'PB2231 (Ser)']
NS2_131 (Thr)	727	['NP59 (Glu)', 'NA447 (Glu)', 'PA275 (Pro)', 'NA313 (Leu)', 'PA186 (Gly)', 'NP322 (Ile)', 'PA85 (Thr)', 'PB1364 (Leu)', 'M1121 (Ala)', 'HA280 (Ile)']
M2_86 (Ala)	726	['M1115 (Ile)', 'NP319 (Tyr)', 'NP289 (Pro)', 'NP67 (Leu)', 'M1137 (Ala)', 'NP295 (Tyr)', 'NP220 (Lys)', 'NP196 (Val)', 'PA341 (Ser)', 'M1121 (Ala)']
PB1_640 (Glu)	726	['PB2653 (Met)', 'PB2205 (Ser)', 'NP196 (Val)', 'NP295 (Tyr)', 'NP319 (Tyr)', 'NP67 (Leu)', 'PB271 (Glu)', 'NP289 (Pro)', 'PB2599 (Gln)', 'PA341 (Ser)']
M2_93 (Ser)	726	['M1115 (Ile)', 'NP319 (Tyr)', 'NP289 (Pro)', 'NP67 (Leu)', 'NP295 (Tyr)', 'M1137 (Ala)', 'NP196 (Val)', 'M1121 (Ala)', 'PA341 (Ser)', 'NP220 (Lys)']

Top 25 Edges

Edges with highest weight are significant since they represent residues with strongest correlated mutations. We list the top 25 edges (with highest MIC values) for the seven datasets in Table 15, Table 16, Table 17, Table 18, Table 19 and Table 20. From these results, we can make the following observations.

The average MIC value for top 25 edges in Avian H5 and H7N9 is about 20% lower compared to averages in H1N1 and H3N2 datasets.

The top edge in Avian H5 dataset (between NS1_214 and NS2_56) has a significantly higher MIC score of 0.95 compared to the next strongest edge (MIC ~0.74).

Majority of the top edges are correlations between residues in non-surface proteins (HA and NA).

All the top edges in Swine H3N2 dataset are edges between M1 and M2 residues.

We see edges with HA or NA residues in Human H1N1 and H3N2 datasets whereas we do not see any HA or NA residues in top 25 edges in Swine H1N1 or H3N2 datasets.

Table 15 - Top 25 edges for Swine H1N1 dataset

SOURCE	TARGET	MIC
NS1_205 (Asn)	NS2_48 (Thr)	0.949
PB1_779 (Ser)	PB2_271 (Ala)	0.912
M1_166 (Ala)	M2_77 (Gln)	0.905
PB1_779 (Ser)	PB2_147 (Thr)	0.904
M1_116 (Ser)	M2_77 (Gln)	0.897
PB1_649 (Asp)	PB2_147 (Thr)	0.894
PB1_649 (Asp)	PB2_271 (Ala)	0.888
M1_101 (Lys)	M2_77 (Gln)	0.886
NS1_194 (Val)	NS2_37 (Ser)	0.88
M1_166 (Ala)	M2_31 (Asn)	0.869
PB1_350 (Met)	PB2_271 (Ala)	0.869
M1_116 (Ser)	M2_31 (Asn)	0.868
PB1_779 (Ser)	PB2_591 (Arg)	0.859
M1_121 (Thr)	M2_77 (Gln)	0.857
PB1_350 (Met)	PB2_147 (Thr)	0.855
NP_386 (Asn)	PB1_779 (Ser)	0.854
PB1_779 (Ser)	PA_400 (Pro)	0.854
M1_101 (Lys)	M2_31 (Asn)	0.849
PB2_147 (Thr)	PA_400 (Pro)	0.849
M1_214 (His)	M2_77 (Gln)	0.848
NP_386 (Asn)	PB2_271 (Ala)	0.845
PB1_649 (Asp)	PA_400 (Pro)	0.842
PB2_271 (Ala)	PA_400 (Pro)	0.842
PB1_649 (Asp)	PB2_591 (Arg)	0.84
NP_386 (Asn)	PB2_147 (Thr)	0.839

Table 16 - Top 25 edges for Human H1N1 Dataset

SOURCE	TARGET	MIC
NS1_205 (Ser)	NS2_48 (Ala)	0.965
M1_80 (Ile)	PA_321 (Lys)	0.943
HA_202 (Thr)	PA_321 (Lys)	0.92
HA_202 (Thr)	M1_80 (Ile)	0.908
NA_321 (Val)	PA_362 (Lys)	0.895
NA_34 (Ile)	PA_362 (Lys)	0.891
NA_432 (Lys)	PA_362 (Lys)	0.887
NA_44 (Asn)	PB1_397 (Ile)	0.879

HA_471 (Asn)	PA_321 (Lys)	0.871
M1_230 (Lys)	PB2_731 (Val)	0.869
PB1_154 (Gly)	PB2_293 (Arg)	0.869
M1_230 (Lys)	PB2_66 (Met)	0.859
M1_230 (Lys)	PB2_293 (Arg)	0.856
HA_471 (Asn)	M1_80 (Ile)	0.854
HA_233 (Ile)	NS1_125 (Glu)	0.852
M1_192 (Met)	PB2_66 (Met)	0.852
NA_264 (Val)	NS2_83 (Met)	0.851
NS1_125 (Glu)	NS2_83 (Met)	0.848
NA_200 (Asn)	M1_230 (Lys)	0.846
NA_44 (Asn)	NS1_90 (Leu)	0.843
PB2_344 (Met)	PA_321 (Lys)	0.843
PB1_397 (Ile)	NS1_90 (Leu)	0.836
HA_13 (Ala)	NS2_83 (Met)	0.834
NA_321 (Val)	PA_100 (Val)	0.834
NA_34 (Ile)	PA_100 (Val)	0.832

Table 17 - Top 25 edges for Swine H3N2 Dataset

SOURCE	TARGET	MIC
M1_116 (Ser)	M2_77 (Gln)	0.931
M1_121 (Thr)	M2_77 (Gln)	0.92
M1_214 (His)	M2_77 (Gln)	0.92
M1_101 (Lys)	M2_77 (Gln)	0.91
M1_166 (Ala)	M2_77 (Gln)	0.908
M1_121 (Thr)	M2_20 (Ser)	0.907
M1_214 (His)	M2_20 (Ser)	0.907
M1_116 (Ser)	M2_20 (Ser)	0.897
M1_95 (Arg)	M2_77 (Gln)	0.881
M1_101 (Lys)	M2_20 (Ser)	0.879
M1_166 (Ala)	M2_20 (Ser)	0.879
M1_95 (Arg)	M2_20 (Ser)	0.87
M1_181 (Leu)	M2_60 (Lys)	0.859
M1_142 (Ala)	M2_77 (Gln)	0.845
M1_142 (Ala)	M2_20 (Ser)	0.838
M1_209 (Thr)	M2_77 (Gln)	0.831
M1_116 (Ser)	M2_31 (Asn)	0.826

M1_209 (Thr)	M2_20 (Ser)	0.824
M1_15 (Ile)	M2_77 (Gln)	0.815
M1_121 (Thr)	M2_31 (Asn)	0.813
M1_214 (His)	M2_31 (Asn)	0.813
M1_30 (Ser)	M2_43 (Thr)	0.808
M1_121 (Thr)	M2_95 (Glu)	0.807
M1_214 (His)	M2_95 (Glu)	0.807
M1_166 (Ala)	M2_31 (Asn)	0.807

Table 18 - Top 25 edges for Human H3N2 Dataset

SOURCE	TARGET	MIC
NA_31 (Phe)	NP_406 (Thr)	0.962
NP_406 (Thr)	PB1_709 (Ile)	0.962
NA_26 (Ser)	NP_406 (Thr)	0.958
NA_50 (Phe)	NP_406 (Thr)	0.95
NA_50 (Phe)	PB1_709 (Ile)	0.945
NA_31 (Phe)	PB1_709 (Ile)	0.942
NA_26 (Ser)	PB1_709 (Ile)	0.94
NA_394 (Asn)	NP_406 (Thr)	0.938
NA_394 (Asn)	PB1_709 (Ile)	0.934
NP_406 (Thr)	PB2_590 (Ser)	0.93
PB1_709 (Ile)	PB2_590 (Ser)	0.927
NA_394 (Asn)	PB2_590 (Ser)	0.92
NA_50 (Phe)	PB2_590 (Ser)	0.917
NP_136 (Ile)	PB1_709 (Ile)	0.916
NA_31 (Phe)	PB2_590 (Ser)	0.913
NA_26 (Ser)	PB2_590 (Ser)	0.91
NA_50 (Phe)	NP_136 (Ile)	0.906
HA_246 (Pro)	PB1_709 (Ile)	0.904
HA_246 (Pro)	NP_406 (Thr)	0.902
NA_31 (Phe)	NP_136 (Ile)	0.901
NA_26 (Ser)	PB1_619 (Asn)	0.9
NP_136 (Ile)	PB2_590 (Ser)	0.9
NA_38 (Ile)	NP_406 (Thr)	0.897
NA_316 (Ile)	NP_406 (Thr)	0.896
HA_246 (Pro)	NA_50 (Phe)	0.893

Table 19 - Top 25 edges for Avian H5 dataset

SOURCE	TARGET	MIC
NS1_214 (Leu)	NS2_56 (Phe)	0.95
M1_232 (Asn)	M2_14 (Glu)	0.739
NS1_214 (Leu)	NS2_22 (Ala)	0.739
HA_228 (Leu)	M1_224 (Asn)	0.727
NP_377 (Asn)	NS1_84 (Met)	0.726
HA_126 (Ile)	M1_166 (Ala)	0.719
HA_126 (Ile)	M1_224 (Asn)	0.718
HA_349 (Lys)	M1_166 (Ala)	0.718
NS1_84 (Met)	NS2_22 (Ala)	0.717
HA_126 (Ile)	M1_232 (Asn)	0.715
NS1_219 (Lys)	NS2_22 (Ala)	0.713
HA_125 (Arg)	M1_232 (Asn)	0.712
HA_125 (Arg)	M1_166 (Ala)	0.708
HA_236 (Ser)	M1_166 (Ala)	0.704
HA_236 (Ser)	M1_224 (Asn)	0.702
HA_418 (Arg)	M1_232 (Asn)	0.702
NS1_219 (Lys)	NS2_56 (Phe)	0.7
NS1_84 (Met)	NS2_56 (Phe)	0.699
HA_418 (Arg)	M1_166 (Ala)	0.697
HA_125 (Arg)	M1_224 (Asn)	0.695
HA_236 (Ser)	M1_232 (Asn)	0.692
HA_126 (Ile)	M1_15 (Ile)	0.688
NP_483 (Asn)	PB1_14 (Val)	0.687
HA_418 (Arg)	M1_224 (Asn)	0.684
NS1_219 (Lys)	NS2_61 (Ile)	0.683

Table 20 - Top 25 edges for H7N9 dataset

SOURCE	TARGET	MIC
NP_375 (Asp)	PB1_397 (Ile)	0.751
PB1_694 (Asn)	PB2_139 (Val)	0.727
NP_375 (Asp)	PB1_171 (Met)	0.719
PB1_397 (Ile)	NS1_152 (Glu)	0.714
NA_83 (Arg)	PB1_397 (Ile)	0.654
PB1_694 (Asn)	PB2_676 (Met)	0.654
PB1_171 (Met)	NS1_152 (Glu)	0.649
PB1_397 (Ile)	NS1_216 (Thr)	0.634

PB1_397 (Ile)	NS1_111 (Ile)	0.628
PB1_397 (Ile)	NS1_80 (Ser)	0.628
PB1_397 (Ile)	NS1_27 (Leu)	0.624
NP_371 (Met)	PB1_397 (Ile)	0.622
NA_83 (Arg)	NP_375 (Asp)	0.617
NP_375 (Asp)	NS1_152 (Glu)	0.61
NA_83 (Arg)	PB1_171 (Met)	0.606
NA_327 (Asn)	PB2_647 (Ile)	0.594
PB1_397 (Ile)	PB2_139 (Val)	0.594
PB1_397 (Ile)	NS1_212 (Ser)	0.591
PB1_171 (Met)	PB2_139 (Val)	0.589
PB1_171 (Met)	NS1_27 (Leu)	0.589
NA_327 (Asn)	PB2_535 (Met)	0.582
NP_375 (Asp)	NS1_111 (Ile)	0.579
NA_327 (Asn)	PB2_511 (Val)	0.576
NP_371 (Met)	PB1_171 (Met)	0.576
PB1_171 (Met)	NS1_216 (Thr)	0.573

Entropy

To gain a global view of sequence variation, we calculated average entropy values for sequences of all the 10 proteins in different datasets (Figure 65). This average entropy plot revealed that NA protein in Avian H5 has the highest overall sequence variation. Proteins in Human-All, Avian H5, Swine H1N1 and Swine H3N2 datasets had higher entropies compared to proteins in H7N9, Human H1N1 and Human H3N2 data sets. Within each dataset, HA, NA and NS1 had the highest average entropy among all the proteins.

We have also created separate plots of average entropies for in-network and out-of-network residues (Figure 66, Figure 67). These plots revealed that the entropy values of in-network residues are generally higher than the entropy values of out-of-network residues.

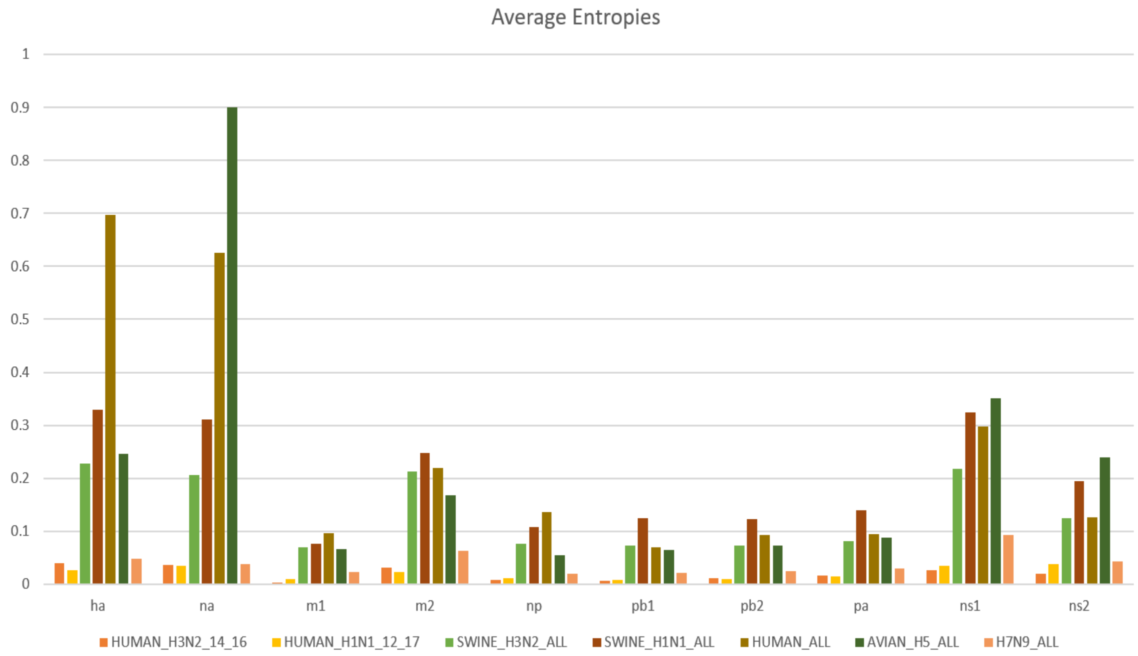


Figure 65 - Average Entropies

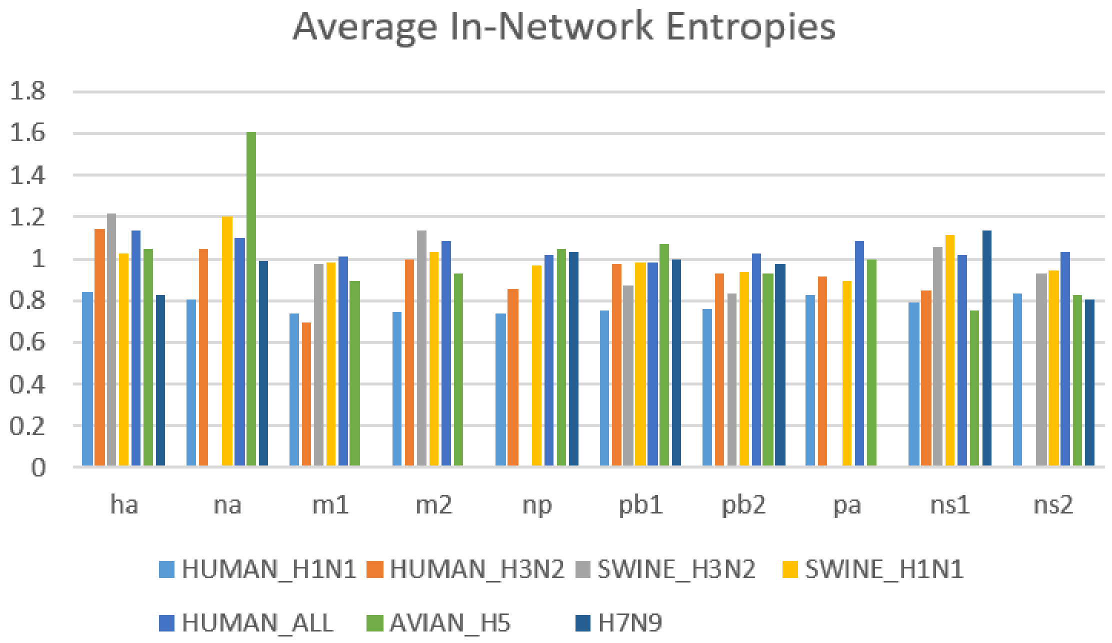


Figure 66 - Average Entropies for In-Network Residues

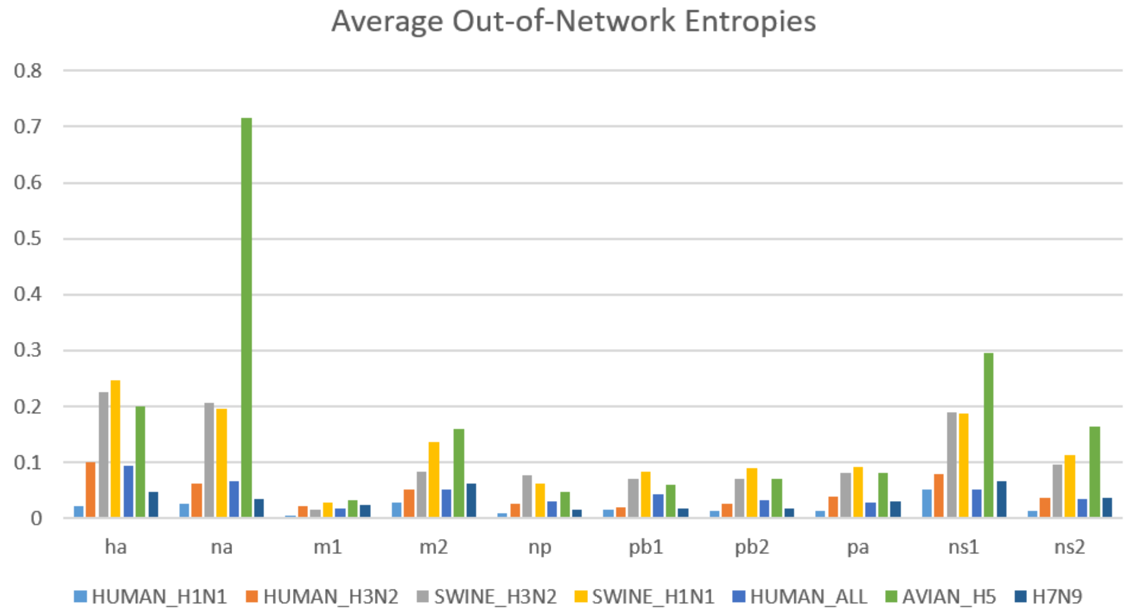


Figure 67 - Average Entropies for Out-of-Network Residues

Solvent Accessibility

We hypothesized that residues with higher solvent accessibility values tend to have a higher probability to participate in inter-protein contacts and thereby have a greater likelihood to be in the network. We calculated solvent accessibility values based on ACC scores in PDB structures for in-network and out-of-network residues in HA, NA, M1, NS1 and NP proteins using 1RU7, 3BEQ, 1EA3, 2GX9 and 2IQH PDB structures respectively. The average values for in-network residues were not statistically higher for most proteins in several datasets compared to the average values for out-of-network cases with few exceptions. The average values for in-network residues in NS1 were statistically higher (2 sample t-test, $p<0.05$) compared to out-of-network residues in all datasets except SWINE_H3N2.

These results disprove our hypothesis and suggest that there is participation of both surface as well as buried residues in networks of correlated mutations.

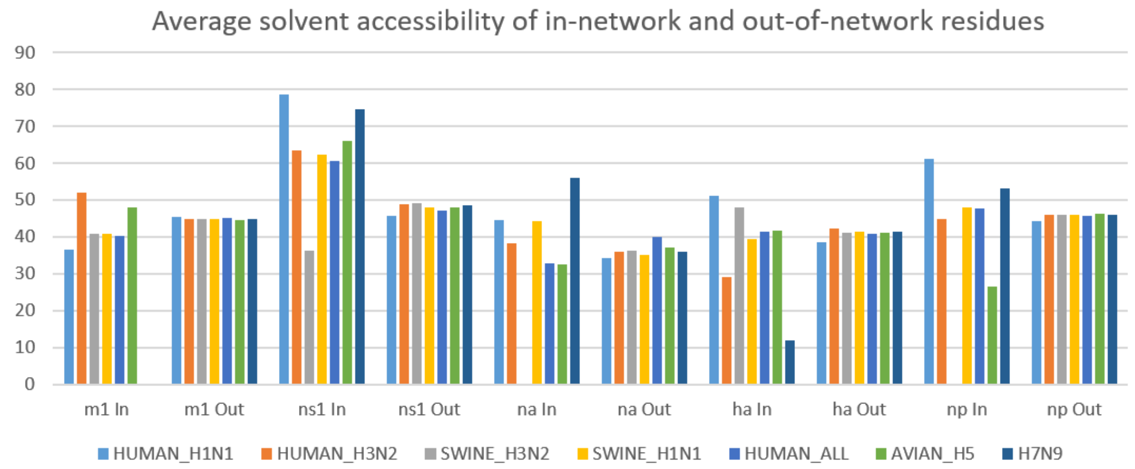


Figure 68 - Average solvent accessibility for in-network and out-of-network residues

Residue Cooccurence Counts

To have a deeper understanding of the association between MIC correlation values and mutations occurring between two residues, we computed residue cooccurrence counts for edges with different MIC values. From these counts, we can interpret that higher MIC scores are generally related to higher degree of covariance in residue pairs. If there is mutation in only one residue, it does not result in a higher score. As an example, the ha_6, ns2_34 edge in Human H1N1 has a weight of only 0.5 since there are proportionately fewer covarying mutations compared to the ns2_48, ns1_205 edge with a weight of 0.965. These counts substantiate the statistical strength of the MIC methodology and provide additional context to our results.

Table 21 - Residue Cooccurence Counts

Dataset	Edge	Weight	Cooccurence Counts
HUMAN_H1N1	ns2_48, ns1_205	0.965	'AS' - 972; 'TN' - 583; 'AN' - 238; 'TS' - 238; 'SS' - 18; 'AI' - 18; 'A-' - 1; 'AR' - 1; 'SI' - 1; 'TK' - 1; 'NN' - 1
HUMAN H1N1	pb2 731, m2 21	0.805	'VD' - 808; 'IG' - 672; 'ID' - 394; 'VG' - 260; 'VV' - 146; '-G' - 1
HUMAN H1N1	ha_6, ns2_34	0.5	'VR' - 1042; 'LR' - 489; 'VQ' - 488; 'AR' - 33; 'IQ' - 4; 'MR' - 4; '-R' - 3; 'LQ' - 3; '-Q' - 2; 'AQ' - 1; 'FR' - 1; 'XQ' - 1; 'LL' - 1
HUMAN H3N2	np_406, na_31	0.962	'TF' - 1003; 'IF' - 584; 'TL' - 568; 'IL' - 420; '-L' - 2; 'TV' - 1; 'XL' - 1; 'I-' - 1; 'IS' - 1
HUMAN H3N2	pb1_586, ha_244	0.5	'RG' - 929; 'KD' - 604; 'RN' - 540; 'KN' - 276; 'RD' - 144; 'KG' - 53; 'K-' - 5; 'RB' - 3; 'XG' - 1; 'KB' - 1
SWINE H1N1	'm1_116', 'm2_77'	0.931	'SQ' - 501; 'AR' - 280; 'SR' - 7; 'AQ' - 6
SWINE H1N1	'm1_142', 'm2_31'	0.771	'AN' - 491; 'VS' - 265; 'VN' - 31; 'AS' - 6; 'VG' - 1
SWINE H1N1	'm1_116' - 'm2_14'	0.5	'SE' - 456; 'AG' - 256; 'SG' - 52; 'AE' - 30
SWINE H3N2	'ns1_205', 'ns2_48'	0.949	'NA' - 384; 'ST' - 364; 'NT' - 144; 'SA' - 48; 'KT' - 35; 'NS' - 27; 'IT' - 18; 'SN' - 14; 'KN' - 12; 'NN' - 12; 'GT' - 10; 'DT' - 6; 'SX' - 3; 'XA' - 3; 'GA' - 3; 'IA' - 2; 'NV' - 2; 'SS' - 2; 'DA' - 1; 'KA' - 1; 'VS' - 1; '-A' - 1; 'RN' - 1; 'ND' - 1; 'S-' - 1
SWINE H3N2	'm2_95', 'm1_214'	0.796	'EQ' - 394; 'VH' - 329; 'EH' - 311; 'AH' - 46; 'MH' - 5; '-H' - 2; 'VQ' - 2; 'SH' - 2; '--' - 1; 'EN' - 1; 'VN' - 1; '-Q' - 1; 'EY' - 1
SWINE H3N2	'ns1_129', 'na_454'	0.5	'IA' - 435; 'TV' - 232; 'VV' - 231; 'IV' - 108; 'IT' - 61; 'TA' - 9; 'VA' - 8; 'TT' - 4; 'AV' - 3; 'IM' - 2; 'VG' - 1; 'LV' - 1; 'MV' - 1
AVIAN H5	'ns2_56', 'ns1_214'	0.95	'FL' - 674; 'LP' - 592; 'LL' - 94; 'FP' - 86; 'LS' - 10; 'FS' - 3; 'CP' - 2; 'LA' - 1; 'FF' - 1
AVIAN H5	'ns1_94', 'ns2_64'	0.5	'TG' - 1266; 'SA' - 178; 'TE' - 7; 'SG' - 4; 'TA' - 4; 'AG' - 1; 'TR' - 1; 'NG' - 1; 'NA' - 1
H7N9	'pb1_397', 'np_375'	0.751	'ID' - 227; 'ME' - 144; 'IE' - 33; 'MD' - 28; 'MN' - 1; 'IG' - 1
H7N9	'ns2_27' - 'na_83'	0.5	'DR' - 234; 'DK' - 94; 'GK' - 55; 'GR' - 50; 'GN' - 1

Protein Correlation Graphs

To understand the extent of co-variation between the 10 proteins in IAV sequences, we created visualizations of protein correlation graphs where the proteins acts as nodes and connections between these nodes are derived based on correlated mutations between residues. Strength of a node (enumerated in parenthesis as part of the name of the node) is the total number of residues in that protein with at least one significant correlation with a residue in another protein, while the strength of a connection (depicted by the thickness of a connection) is the total number of edges between residues in two proteins. Several interesting observations can be made from these visualizations (Figure 69, Figure 70, Figure 71, Figure 72, Figure 73, Figure 74 and Figure 75).

1. HA and NA proteins play the most prominent role in protein correlation networks. They tend to have the maximum number of residues with significant correlations.
2. NA protein dominates the Avian H5 network (Figure 73).
3. Figure 71 and Figure 72 elucidate the differences between Human H3N2 and Swine H3N2 networks. The Swine H3N2 network is sparse and contains very few connections between proteins compared to the Human H3N2 network.
4. Protein interaction networks of Human H1N1 and Human H3N2 (Figure 70, Figure 73) suggest that residues in NP have the third highest number of residues with correlated mutations (after HA and NA).
5. These networks (except for H7N9 network) contain residues from all 10 proteins.

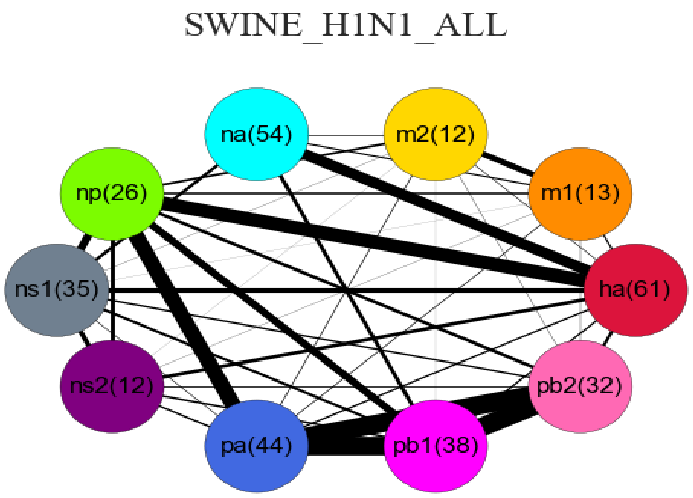


Figure 69 - Protein Correlation Network for Swine H1N1 Dataset

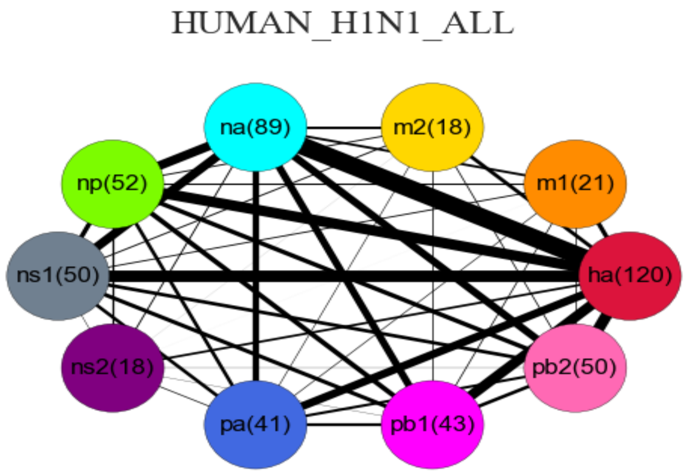


Figure 70 - Protein Correlation Network for Human H1N1 Dataset

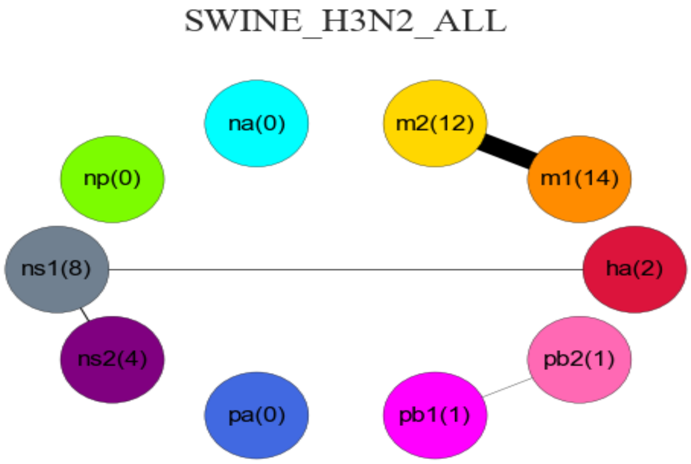


Figure 71 - Protein Correlation Network for Swine H3N2 Dataset

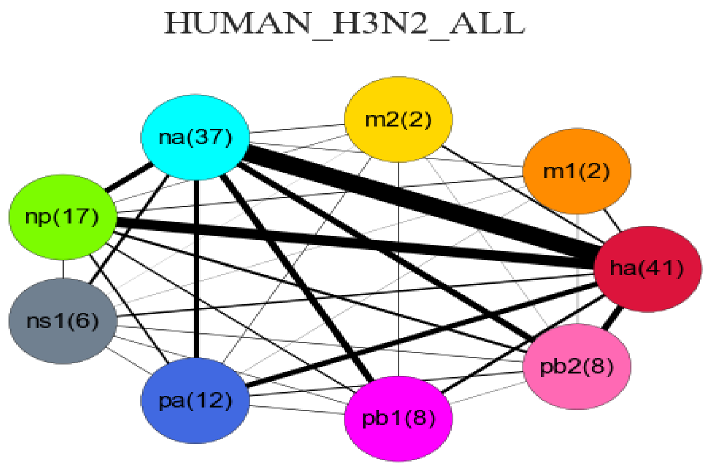


Figure 72 - Protein Correlation Network for Human H3N2 Dataset

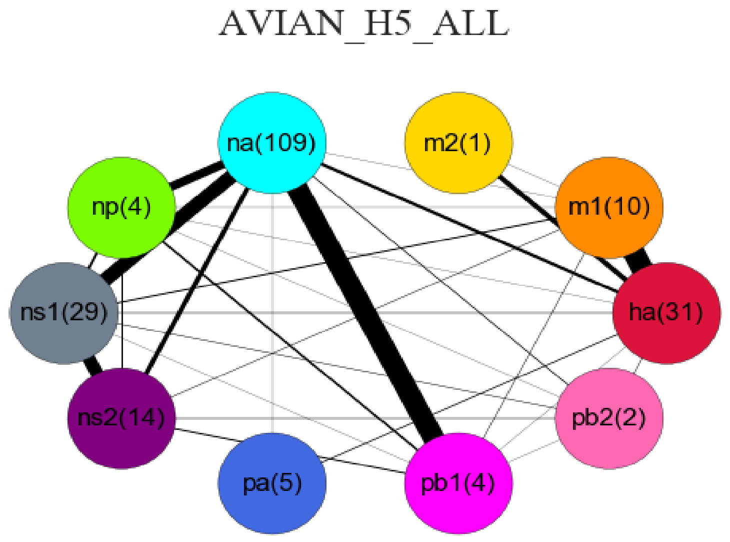


Figure 73 - Protein Correlation Network for Avian H5 Dataset

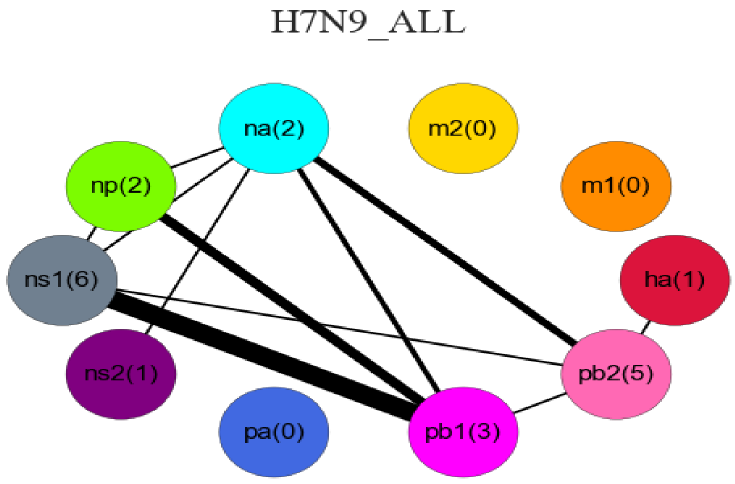


Figure 74 - Protein Correlation Network for H7N9 Dataset

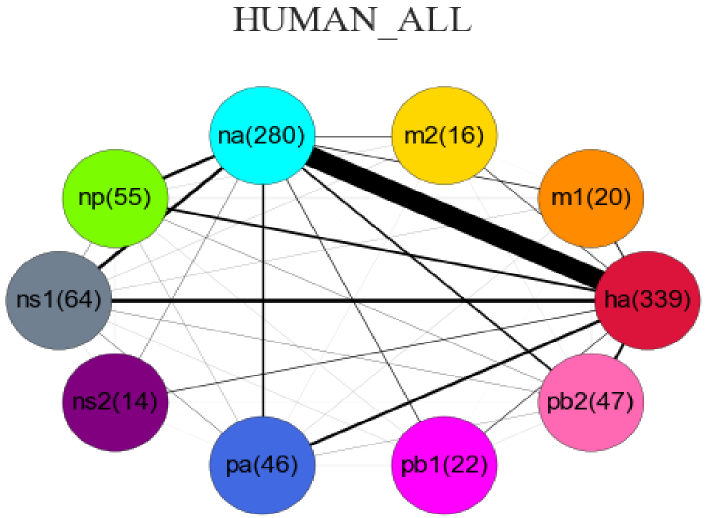


Figure 75 - Protein Correlation Network for Human_All Dataset

Evolution of Human H1N1 Network

To visualize how H1N1 viruses in Humans evolve over time, we examined the changes in network topology over time by analyzing four different datasets listed in Table 22. We started with a dataset of 698 unique strains belonging to years till 2007. For our second dataset, we created a dataset consisting of 1035 unique strains belonging to flu seasons till 2010/2011 with a maximum of 300 strains per year. Node and edge counts for these datasets are depicted in Figure 76 and Figure 77 respectively.

These results suggest that there is a sharp increase in epistatic changes in Influenza genome between 2007 and 2010/2011 flu season during which a pandemic H1N1 occurred. There is a gradual normalization in these changes in subsequent years where the genome continued to see lesser covariance in mutations. Both the node and edge count observations confirm this hypothesis. If the trend continues (where we do not observe any new H1N1 pandemics soon), we should continue to see a similar pattern with the long-tail in the number of nodes and a gradual decrease in number of edges in the coming years.

Table 22 - Datasets for studying Human H1N1 evolution

NAME	#STRAINS	COMMENTS
HUMAN_H1N1_TILL_2007	698	698 unique strains of H1N1 belonging to years till 2007
HUMAN_H1N1_TILL_10_11	1035	1035 unique strains of H1N1 belonging to flu seasons before 2010/11, maximum 300 per year
HUMAN_H1N1_TILL_14_15	1448	1448 unique strains of H1N1 belonging to flu seasons before 2014/15, maximum 300 per year
HUMAN_H1N1_TILL_16_17	1769	1769 unique strains of H1N1 belonging to flu seasons before 2016/17, maximum 300 per year

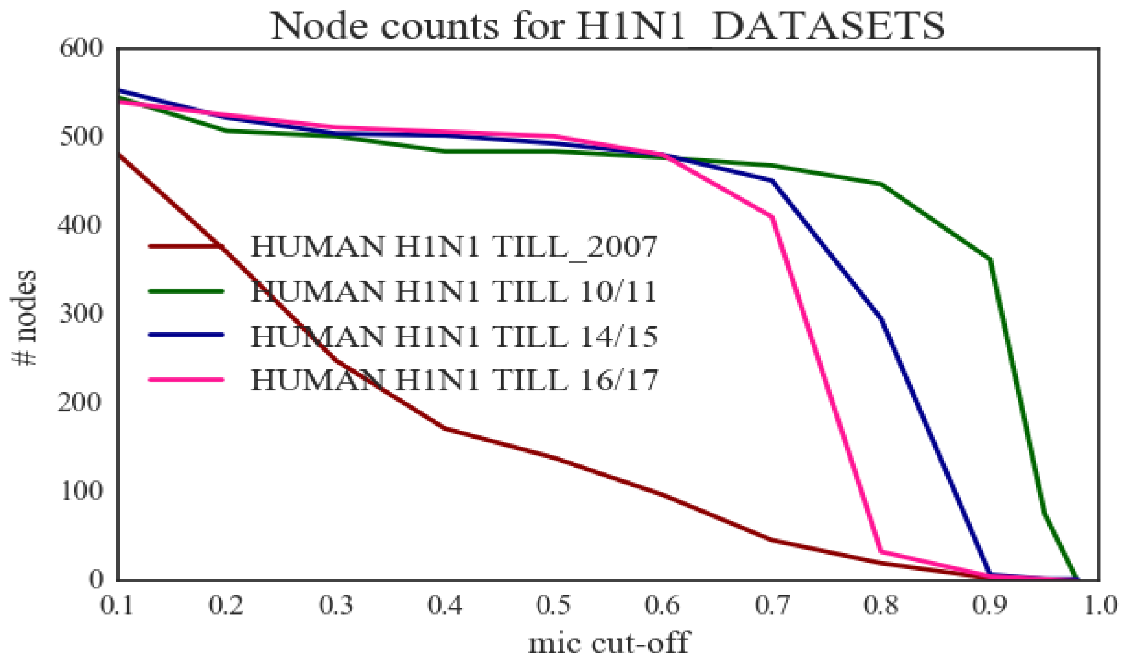


Figure 76 - Node counts for Human H1N1 Datasets

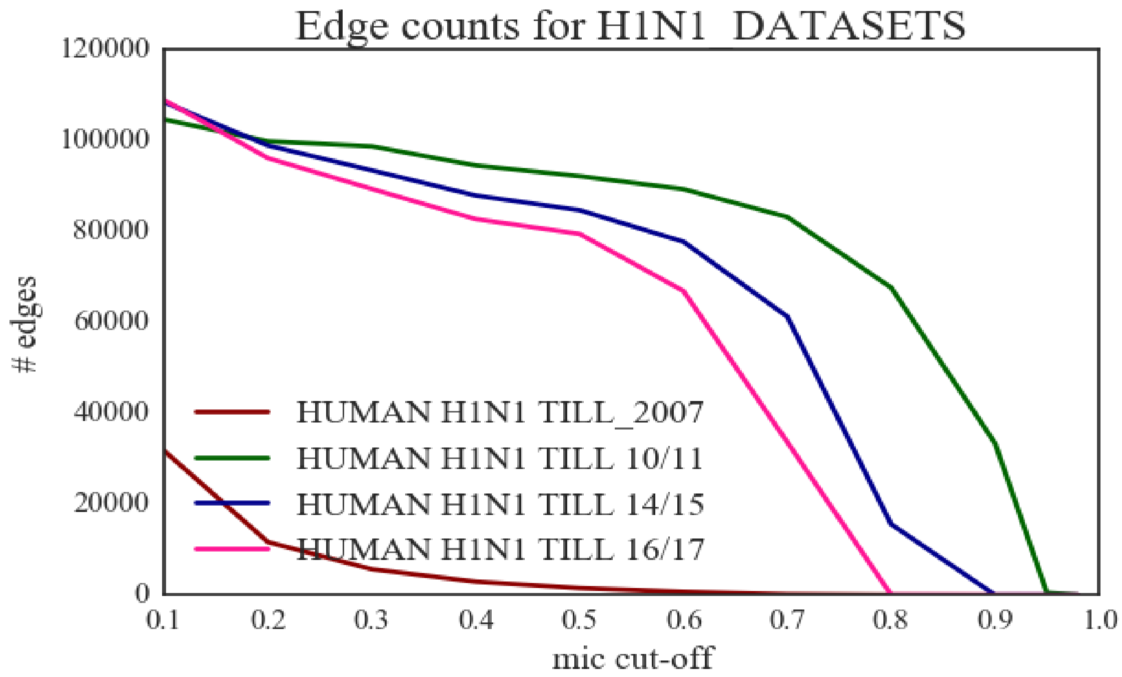


Figure 77 - Edge counts for Human H1N1 Datasets

Evolution of Human H3N2 Network

To visualize how H3N2 viruses in Humans evolve over time, we examined the changes in network topology over time by analyzing four different datasets over an increasing 10-year window (listed in Table 23). Compared to the H1N1 evolution plots, the H3N2 plots show a far lesser degree of change over the last 10 years suggesting a more stable, less dynamic genome. Also, the number of nodes and edges in the H3N2 network are substantially lesser than the number of nodes and edges in H1N1 network.

Table 23 - Datasets for studying Human H3N2 evolution

NAME	#STRAINS	COMMENTS
H3N2_HUMAN_TILL_2016	1561	Unique - HUMAN H3N2 - BEFORE 2007 - MAX 300 per year

H3N2_HUMAN_TILL_2010	1913	Unique - HUMAN H3N2 - BEFORE 2010 - MAX 200 per year
H3N2_HUMAN_TILL_2014	2154	Unique - HUMAN H3N2 - BEFORE 2014 - MAX 150 per year
H3N1_HUMAN_ALL	1940	Unique - HUMAN H3N2 – ALL – MAX 100 per year

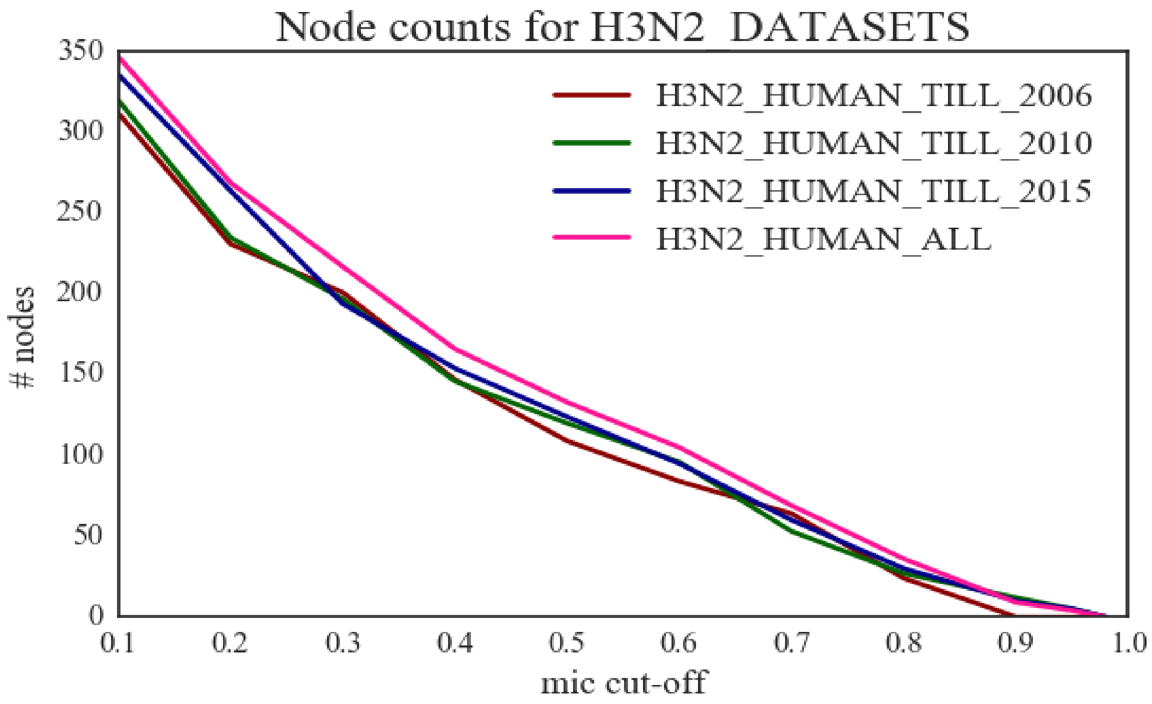


Figure 78 - Node counts for Human H3N2 Datasets

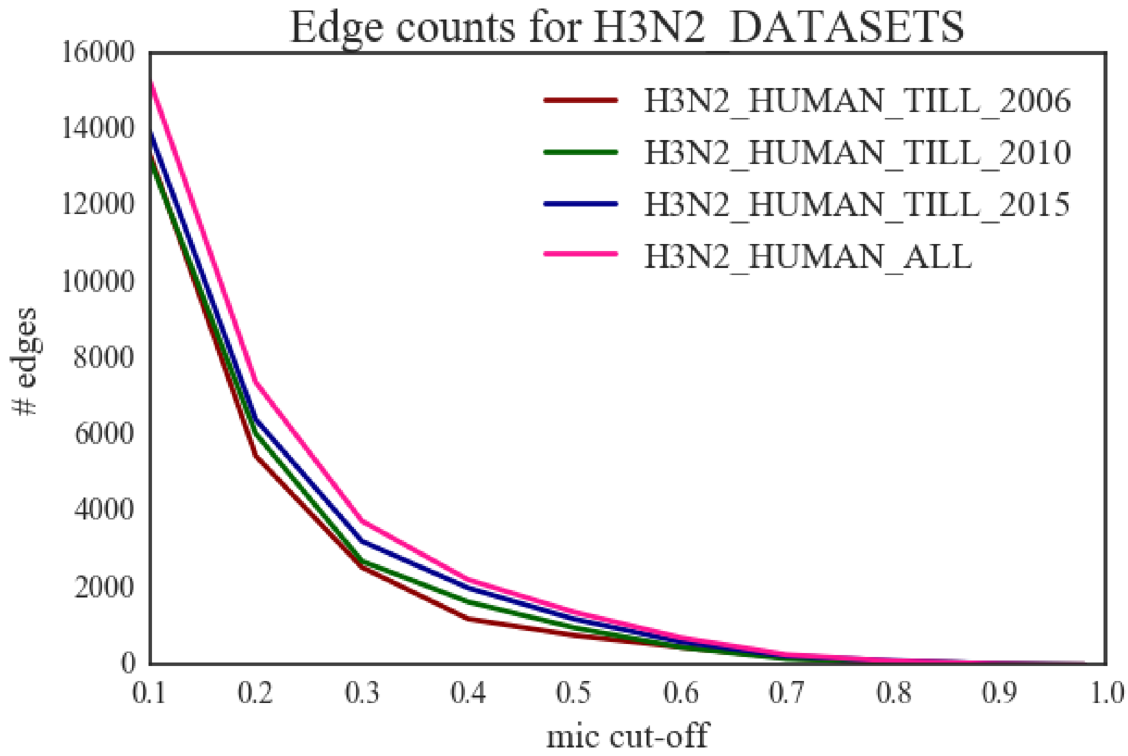


Figure 79 - Edge counts for Human H3N2 Datasets

CHAPTER 9 – CONCLUSIONS

In this study, we conceived and realized a pipeline for systematic analysis of the structural dynamics of IAV system using a network approach. We employed a non-linear correlation measure called 'Maximal Information Coefficient' (MIC) to identify correlated mutations and created networks based on MIC scores. We created separate networks for seven different primary datasets – HUMAN H1N1, SWINE H1N1, HUMAN H3N2, SWINE H3N2, AVIAN H5, H7N9 and HUMAN ALL and studied topological properties of the networks. In a separate but related effort, we successfully applied 'normalized n-gram frequencies' as feature vectors to classify IAV sequences. Our study led us to several important conclusions.

Network Analysis of Correlated Mutations in IAV can provide novel insights

Based on our study, we can conclude that correlated mutation networks provide a unique and new perspective to understanding the structural dynamics of IAV system. These networks can supplement and extend existing structure based computational approaches and distance based network models to augment our understanding of the overall features of viral proteome. Our approach provides a robust framework to quantify correlated mutations, evaluate the mutational space of each amino acid and visualize the impact of a specific mutation in a IAV protein on amino acid residues in other proteins. We could identify residues that act as hubs and find edges and triplets with maximum covariance that can be useful for epitope identification and antibody engineering.

Correlated Mutation Networks in IAV are sub-type and host specific

By implementing our pipeline on datasets covering different sub-types and hosts, we can conclude that the correlated mutation networks are both sub-type and host specific. There are substantial differences in the networks of Swine H1N1 and Human H1N1 network and similarly between networks of Swine H3N2 and Human H3N2 networks. There are more pronounced differences between networks of Human H1N1 and Human H3N2 networks (and similarly between Human H3N2 and Swine H3N2 networks. We have also analyzed the correlated mutation networks of Avian H5 and H7N9 and found them to be significantly different from other networks. The correlated mutation network created from a combination of strains of different sub-types turned out to be the densest and was significantly different from all other networks.

Correlated Mutation Networks can improve our understanding of Influenza evolution

Our experiments with Human H1N1 and H3N2 datasets over an increasing time allowed us to gain a unique perspective into evolution of H1N1 and H3N2 in Humans based on correlated mutation profiles. We observed that H3N2 strains in Humans followed a stable pattern where we saw a slow and gradually growing network over the years. We observed a sudden increase in the number of nodes and edges in The H1N1 network from 2006/2007 flu season to the 2009/2010 flu season and the network showed signs of stability and slow, gradual changes after the 2009/2010 season.

Observations from Correlated Mutation Networks in Influenza (@ 0.5 threshold)

We could make several interesting observations based on our results. First, Nodes with highest degree and edges with strongest weight did not generally belong to virally active surface proteins (HA and NA) for all datasets although a large percentage (>50%) of in-network residues belong to these two proteins in Human H1N1, Swine H1N1, Human H3N2 and Avian H5 networks. The Avian H5 network is dominated by residues from NA (52%). The H7N9 and Swine H3N2 networks @ 0.5 threshold are very sparse with only 20 and 42 nodes respectively.

There is a statistically significant correlation between entropy and MIC correlations

Our study clearly highlighted that residues with high degree of entropy have much higher probability to be in-network compared to residues with lesser entropy. While we observed statistical significance in the association between MIC and entropy values, it is important to mention here that entropy does not necessarily imply a high MIC value and we have seen cases where residues with high entropy were not in the network.

There is no statistically significant correlation between solvent accessibility and MIC correlations

We explored potential association between residues with high correlated mutations and their solvent accessibility scores and did not observe and statistical correlation suggesting that both surface and buried residues are involved in correlated mutations.

Computational pipeline and software can be applied to other systems

We have developed a robust and comprehensive software pipeline to pre-process data, compute MIC, create MIC-based networks and analyze the networks. This pipeline can be reused to other viral and small non-viral systems with sufficient data.

Normalized 3-gram counts are reliable features for automated classification of IAV sequences

In our ‘classification’ work, we proved that N-gram analysis coupled with supervised classification algorithms to distinguish between strains is a sound approach. We created a software pipeline in Python and applied it to classify protein and nucleotide sequences of IAV that we downloaded from IRD. Using this approach, we could perform binary classification to distinguish sequences from different subtypes, (consecutive) flu seasons, geographic locations (Asia vs North-America), similarity to 2009 pandemic H1N1 and most importantly drug resistance. We could classify sequences of HA, NA, M1, M2 and NA based on their resistance to Adamantane and Oseltamavir with a near 100% accuracy. Using Random Forest classifier, we identified the most significant features (3-grams) in NA and M2 sequences and could confirm a strong linkage between these features and known drug resistant mutations in these proteins.

CHAPTER 10 - FUTURE WORK

This study focused on performing comprehensive analysis of correlated mutations in IAV sequences using a network approach. We could conduct this study primarily because of the growing number of IAV sequences in public repositories. Results of our study can improve our understanding of the evolutionary dynamics of IAV. We have presented several specific conclusions based on the results of our work. Our effort provided a framework for understanding correlated mutations and overall dynamics of a system based on a network approach. We have developed a methodology and a computational pipeline that can serve as an additional tool for in-vitro mutation analysis of similar biological systems using large number of protein sequences.

We see several opportunities for application of our approach to other similar systems and for improvement of our approach.

First and foremost, we can incorporate additional sequences from specific years and/or regions into our existing network to understand the impact. Such efforts can improve the adoption of this approach and can make it a new tool to not only study the dynamics of the system more systematically but also to improve our understanding of epidemiological patterns.

Second, the computational pipeline that we developed can be applied to other biological systems of similar size. The combined sequence length of the 10 proteins in IAV is approximately 4500 and we believe that our pipeline can scale without any issues for systems with up to 20000 residues. Other viral systems like Ebola, West Nile, Chikungunya and Dengue are good candidates for similar study in the future based on availability of sufficient sequences.

There is also scope for improvement of our overall computational approach. While we employed a Boolean [0,1] notation to represent our sequences, we believe that there is value in trying other notations that more accurately capture the strength of a mutation before performing MIC computations. In fact, we have some preliminary results based on the use of blosum62 scoring matrix and a second approach that employs the hydrophobicity of residues as alternate notations that we did not report in this thesis. Comparing the results of networks created based on these alternate notations with a Boolean [0,1] notation can improve the overall quality of results presented in this work. While we have taken some steps to understand the impact of background noise, much work requires to be done in this area to automatically identify mutations that are intrinsically phylogenetic in nature and include such a step as part of the overall pipeline. The other area that is worth pursuing is systematically comparing MIC-based mutation networks with more conventional mutual information derived mutation networks to improve robustness of the overall approach. For the classification work that we reported in this thesis, the increasing number of labeled and curated sequences available in IRD presents a unique opportunity for application of deep learning techniques to perform classification without feature engineering.

CHAPTER 11 - APPENDIX

Intra-Protein Correlated Mutations in IAV

While our work primarily focused on analyzing inter-protein correlated mutations in IAV sub-systems, we have also created intra-protein correlated mutation graphs to understand differences and gain relevant insights. In this appendix, we have included results for intra-protein correlated mutations within residues in proteins in IAV sub-systems for our 6 primary datasets (listed in Data). Figure 80 and Figure 81 depict the node and edge counts for each protein. These figures suggest significant differences between the 6 datasets. There are significantly higher number of nodes and edges in the 'NA' network of Avian_H5. Figure 81 also suggests that the intra-protein correlated mutations graphs for Human H3N2, Swine H3N2 and H7N9 are very sparse with a significantly lower edge to node ratio.

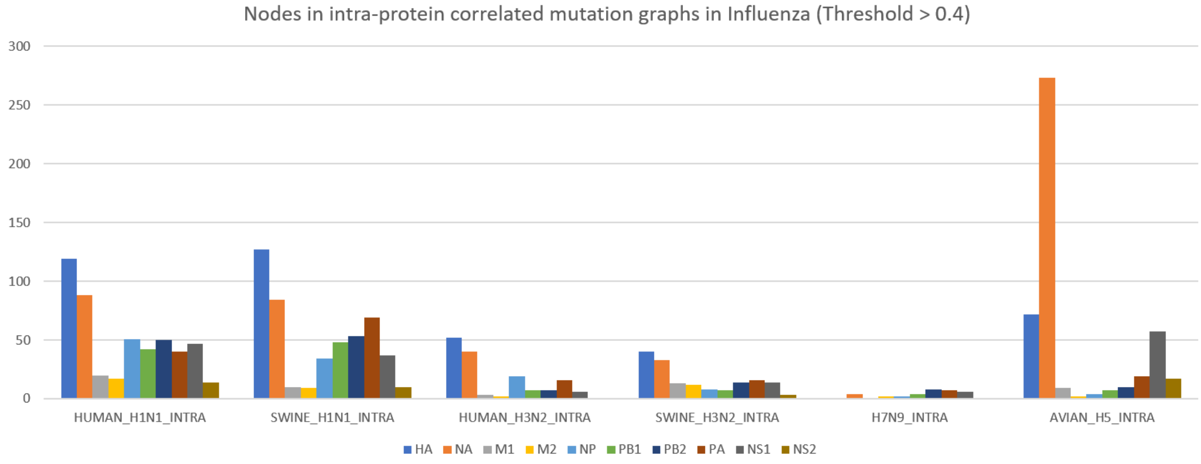


Figure 80 - Node counts in Intra-Protein correlated mutation graphs with MIC threshold set to 0.4

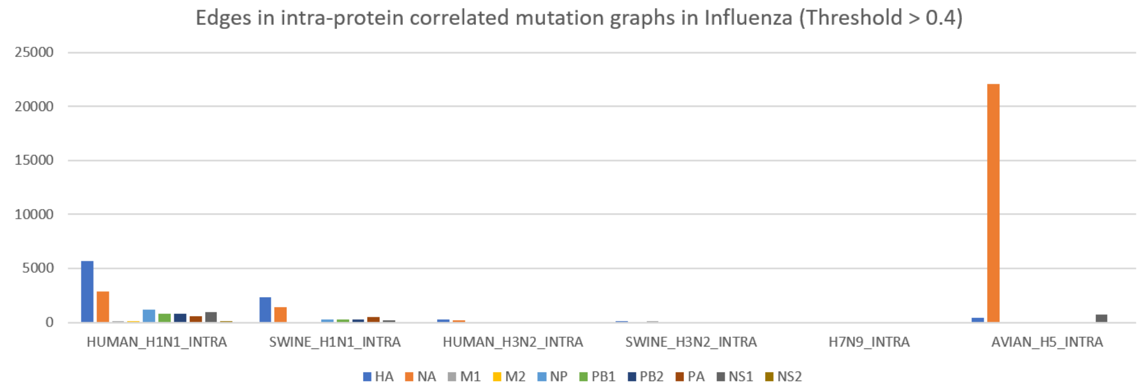


Figure 81 - Edge counts in Intra-Protein correlated mutation graphs with MIC threshold set to 0.4

Source Code

Source code developed as part of this effort can be found in the following github repositories.

- MIC computation, graph creation and analysis of correlated mutations

  https://github.com/yallapragada/network_analysis

- Web application to depict results of network analysis

  https://github.com/yallapragada/network_analysis_web

- Classification of Influenza sequences using ngrams

  https://github.com/yallapragada/ngram_classifier

REFERENCES

[1] E. Ghedin *et al.*, “Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution,” *Nature*, vol. 437, no. 7062, pp. 1162–1166, Oct. 2005.

[2] “WHO | Influenza.” [Online]. Available: http://www.who.int/biologicals/vaccines/influenza/en/. [Accessed: 20-Apr-2015].

[3] D. Altschuh, A. M. Lesk, A. C. Bloomer, and A. Klug, “Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus,” *J. Mol. Biol.*, vol. 193, no. 4, pp. 693–707, Feb. 1987.

[4] I. N. Shindyalov, N. A. Kolchanov, and C. Sander, “Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?,” *Protein Eng. Des. Sel.*, vol. 7, no. 3, pp. 349–358, Mar. 1994.

[5] U. Göbel, C. Sander, R. Schneider, and A. Valencia, “Correlated mutations and residue contacts in proteins,” *Proteins Struct. Funct. Bioinforma.*, vol. 18, no. 4, pp. 309–317, Apr. 1994.

[6] G. B. Gloor, L. C. Martin, L. M. Wahl, and S. D. Dunn, “Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions,” *Biochemistry (Mosc.)*, vol. 44, no. 19, pp. 7156–7165, May 2005.

[7] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan, “Evolutionary information for specifying a protein fold,” *Nature*, vol. 437, no. 7058, pp. 512–518, Sep. 2005.

[8] A. Kowarsch, A. Fuchs, D. Frishman, and P. Pagel, “Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions,” *PLOS Comput. Biol.*, vol. 6, no. 9, pp. 1–13, 2010.

[9] K. Das, J. M. Aramini, L.-C. Ma, R. M. Krug, and E. Arnold, “Structures of influenza A proteins and insights into antiviral drug targets.,” *Nat. Struct. Mol. Biol.*, vol. 17, no. 5, pp. 530–538, May 2010.

[10] K. Viswanathan, Z. Shriver, and G. J. Babcock, “Amino acid interaction networks provide a new lens for therapeutic antibody discovery and anti-viral drug optimization,” *Curr. Opin. Virol.*, vol. 11, no. 0, pp. 122–129, Apr. 2015.

[11] F. Carrat and A. Flahault, “Influenza vaccine: The challenge of antigenic drift,” *Vaccine*, vol. 25, no. 39–40, pp. 6852–6862, Sep. 2007.

[12] P. C. Soema, R. Kompier, J.-P. Amorij, and G. F. A. Kersten, “Current and next generation influenza vaccines: Formulation and production strategies,” *Eur. J. Pharm. Biopharm.*, vol. 94, no. 0, pp. 251–263, Aug. 2015.

[13] J.-M. Song, N. Van Rooijen, J. Bozja, R. W. Compans, and S.-M. Kang, “Vaccination inducing broad and improved cross protection against multiple subtypes of influenza A virus,” *Proc. Natl. Acad. Sci.*, vol. 108, no. 2, pp. 757–761, Jan. 2011.

[14] Y.-N. Lee, M.-C. Kim, Y.-T. Lee, Y.-J. Kim, and S.-M. Kang, “Mechanisms of Cross-protection by Influenza Virus M2-based Vaccines,” *Immune Netw.*, vol. 15, no. 5, p. 213, Oct. 2015.

[15] “Virology Journal | Full text | On the epidemiology of influenza.” .

[16] “Classification of Influenza Viruses.” [Online]. Available: http://www.influenzavirusnet.com/influenza-classification.html. [Accessed: 20-Apr-2015].

[17] “WHO | Updated unified nomenclature system for the highly pathogenic H5N1 avian influenza viruses.” [Online]. Available: http://www.who.int/influenza/gisrs_laboratory/h5n1_nomenclature/en/. [Accessed: 28-Apr-2015].

[18] “Structure of influenza virus.” [Online]. Available: http://www.virology.ws/2009/04/30/structure-of-influenza-virus/. [Accessed: 15-Apr-2015].

[19] “Influenza A life cycle. : Structures of influenza A proteins and insights into antiviral drug targets : Nature Structural & Molecular Biology : Nature Publishing Group.” [Online]. Available: http://www.nature.com/nsmb/journal/v17/n5/fig_tab/nsmb.1779_F1.html. [Accessed: 26-Apr-2015].

[20] T. Horimoto and Y. Kawaoka, “Influenza: lessons from past pandemics, warnings from current incidents,” *Nat Rev Micro*, vol. 3, no. 8, pp. 591–600, Aug. 2005.

[21] “The Viral Genomes Resource.” [Online]. Available: http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239&hopt=scheme. [Accessed: 28-Apr-2015].

[22] E. David and K. Jon, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York, NY, USA: Cambridge University Press, 2010.

[23] B. Bollobas, *Graph Theory - An Introductory Course | Bela Bollobas | Springer*. .

[24] D. West, *Introduction to Graph Theory*, 2 edition. New York, NY: Pearson, 2017.

[25] “CorrelationComparison.pdf.” .

[26] C. D. Rau, N. Wisniewski, L. D. Orozco, B. Bennett, J. Weiss, and A. J. Lusis, “Maximal information component analysis: a novel non-linear network analysis method,” *Front. Genet.*, vol. 4, p. 28, 2013.

[27] D. N. Reshef *et al.*, “Detecting Novel Associations in Large Data Sets,” *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.

[28] Z. Zhang, S. Sun, M. Yi, X. Wu, and Y. Ding, “MIC as an Appropriate Method to Construct the Brain Functional Network,” *BioMed Res. Int.*, vol. 2015, p. 825136, 2015.

[29] R. B. Squires *et al.*, “Influenza research database: an integrated bioinformatics resource for influenza research and surveillance.,” *Influenza Other Respir. Viruses*, vol. 6, no. 6, pp. 404–416, Nov. 2012.

[30] S. Van der Auwera, I. Bulla, M. Ziller, A. Pohlmann, T. Harder, and M. Stanke, “ClassyFlu: Classification of Influenza A Viruses with Discriminatively Trained Profile-HMMs,” *PLoS ONE*, vol. 9, no. 1, p. e84558, 2014.

[31] “The treatment of influenza with antiviral drugs.” [Online]. Available: http://www.cmaj.ca/content/168/1/49.full. [Accessed: 26-Apr-2015].

[32] V. M. Deyde *et al.*, “Surveillance of Resistance to Adamantanes among Influenza A(H3N2) and A(H1N1) Viruses Isolated Worldwide,” *J. Infect. Dis.*, vol. 196, no. 2, pp. 249–257, Jul. 2007.

[33] G. Dong *et al.*, “Adamantane-Resistant Influenza A Viruses in the World (1902–2013): Frequency and Distribution of M2 Gene Mutations,” *PLoS ONE*, vol. 10, no. 3, p. e0119115, Mar. 2015.

[34] “WHO Weekly Pandemic H1N1 Report - 2009, Update 69.”

[35] H. T. Nguyen, A. M. Fry, P. A. Loveless, A. I. Klimov, and L. V. Gubareva, “Recovery of a Multidrug-Resistant Strain of Pandemic Influenza A 2009 (H1N1) Virus Carrying a Dual H275Y/I223R Mutation from a Child after Prolonged Treatment with Oseltamivir,” *Clin. Infect. Dis.*, vol. 51, no. 8, pp. 983–984, Oct. 2010.

[36] Y. Itoh *et al.*, “Emergence of H7N9 Influenza A Virus Resistant to Neuraminidase Inhibitors in Nonhuman Primates,” *Antimicrob. Agents Chemother.*, vol. 59, no. 8, pp. 4962–4973, Aug. 2015.

[37] H.-L. Yen *et al.*, “Resistance to Neuraminidase Inhibitors Conferred by an R292K Mutation in a Human Influenza Virus H7N9 Isolate Can Be Masked by a Mixed R/K Viral Population,” *mBio*, vol. 4, no. 4, pp. e00396-13, 2013.

[38] Z. Volkovich, V. Kirzhner, A. Bolshoy, E. Nevo, and A. Korol, “The method of -grams in large-scale clustering of {DNA} texts,” *Pattern Recognit.*, vol. 38, no. 11, pp. 1902–1912, 2005.

[39] A. Tomović, P. Janičić, and V. Kešelj, “n-Gram-based classification and unsupervised hierarchical clustering of genome sequences,” *Comput. Methods Programs Biomed.*, vol. 81, no. 2, pp. 137–153.

[40] J. K. Vries, X. Liu, and I. Bahar, “The relationship between n-gram patterns and protein secondary structure.,” *Proteins*, vol. 68, no. 4, pp. 830–838, Sep. 2007.

[41] D. Tobi and I. Bahar, “Recruitment of rare 3-grams at functional sites: is this a mechanism for increasing enzyme specificity?,” *BMC Bioinformatics*, vol. 8, p. 226, 2007.

[42] I. Frades, S. Resjo, and E. Andreasson, “Comparison of phosphorylation patterns across eukaryotes by discriminative N-gram analysis.,” *BMC Bioinformatics*, vol. 16, p. 239, 2015.

[43] I. Muhammad, F. Ibrahim, and S. Brahim, “Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics,” *Sci. World J.*, vol. 2014.

[44] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, “Understanding variable importances in forests of randomized trees,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 431–439.

[45] C. D. Rau, N. Wisniewski, L. D. Orozco, B. Bennett, J. N. Weiss, and A. J. Lusis, “Maximal information component analysis: a novel non-linear network analysis method,” *Front. Genet.*, vol. 4, 2013.

[46] R. Ge *et al.*, “McTwo: a two-step feature selection algorithm based on maximal information coefficient,” *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–14, 2016.

[47] P. J. A. Cock *et al.*, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009.

[48] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[49] L. Buitinck *et al.*, “API design for machine learning software: experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[50] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[51] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, “minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers,” *Bioinformatics*, vol. 29, no. 3, pp. 407–408, Feb. 2013.

[52] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, “WebLogo: A Sequence Logo Generator,” *Genome Res.*, vol. 14, no. 6, pp. 1188–1190, Jun. 2004.

[53] J. Wang and W. Wang, “A computational approach to simplifying the protein folding alphabet,” *Nat Struct Mol Biol*, vol. 6, no. 11, pp. 1033–1038, Nov. 1999.

[54] L. R. Murphy, A. Wallqvist, and R. M. Levy, “Simplified amino acid alphabets for protein fold recognition and implications for folding,” *Protein Eng.*, vol. 13, no. 3, pp. 149–152, Mar. 2000.

[55] E. Jacob, R. Unger, and A. Horovitz, “Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis,” *eLife*, vol. 4, p. e08932, Sep. 2015.

[56] P. J. Kundrotas and E. G. Alexov, “Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives,” *BMC Bioinformatics*, vol. 7, pp. 503–503, 2006.

[57] W. R. Taylor, R. S. Hamilton, and M. I. Sadowski, “Prediction of contacts from correlated sequence substitutions,” *New Contructs Expr. Proteins Seq. Topol.*, vol. 23, no. 3, pp. 473–479, Jun. 2013.

[58] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, “PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments,” *Bioinformatics*, vol. 28, no. 2, pp. 184–190, Jan. 2012.

[59] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, “Identification of direct residue contacts in protein–protein interaction by message passing,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 1, pp. 67–72, Jan. 2009.

[60] M. C. Karnoub, F. Seillier-Moiseiwitsch, and P. K. Sen, “A conditional approach to the detection of correlated mutations,” in *Statistics in molecular biology and genetics*, vol. Volume 33, F. Seillier-Moiseiwitsch, Ed. Hayward, CA: Institute of Mathematical Statistics, 1999, pp. 221–235.

[61] M. A. Fares and D. McNally, “CAPS: coevolution analysis using protein sequences,” *Bioinformatics*, vol. 22, no. 22, pp. 2821–2822, Nov. 2006.

[62] Q. Wang and C. Lee, “Distinguishing functional amino acid covariation from background linkage disequilibrium in HIV protease and reverse transcriptase.,” *PloS One*, vol. 2, no. 8, 2007.

[63] Y. Liu, E. Eyal, and I. Bahar, “Analysis of correlated mutations in HIV-1 protease using spectral clustering.,” *Bioinforma. Oxf. Engl.*, vol. 24, no. 10, pp. 1243–1250, May 2008.

[64] O. Haq, R. M. Levy, A. V. Morozov, and M. Andrec, “Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease.,” *BMC Bioinformatics*, vol. 10 Suppl 8, 2009.

[65] W. Mao, C. Kaya, A. Dutta, A. Horovitz, and I. Bahar, “Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution,” *Bioinformatics*, vol. 31, no. 12, pp. 1929–1937, Jun. 2015.

[66] W. Hu, “Analysis of correlated mutations, stalk motifs, and phylogenetic relationship of the 2009 influenza A virus neuraminidase sequences,” *J. Biomed. Sci. Eng.*, vol. 2, pp. 550–558.

[67] W. Hu, “Host markers and correlated mutations in the overlapping genes of influenza viruses: M1, M2; NS1, NS2; and PB1, PB1-F2,” *Nat. Sci.*, vol. 2, pp. 1225–1246, 2010.

[68] W. Hu, “Correlated mutations in the four influenza proteins essential for viral RNA synthesis, host adaptation, and virulence: NP, PA, PB1, and PB2,” *Nat. Sci.*, vol. 2, pp. 1138–1147, 2010.

[69] V. Dahirel *et al.*, “Coordinate linkage of HIV evolution reveals regions of immunological vulnerability.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 28, pp. 11530–11535, Jul. 2011.

[70] F. Tria, S. Pompei, and V. Loreto, “Dynamically correlated mutations drive human Influenza A evolution.,” *Sci. Rep.*, vol. 3, 2013.

[71] M. X. Ruiz-Gonzalez and M. A. Fares, “Coevolution analyses illuminate the dependencies between amino acid sites in the chaperonin system GroES-L.,” *BMC Evol. Biol.*, vol. 13, 2013.

[72] “Amino-Acid Correlated Mutations inside a Single Protein System: A New Method for the Identification of Main Coherent Directions of Evolutive Changes | Open Access | OMICS Publishing Group.” [Online]. Available: http://esciencecentral.org/journals/amino-acid-correlated-mutations-inside-a-single-protein-system-a-new-method-for-the-identification-of-main-coherent-directions-of-evolutive-changes-2329-9002.1000111.php?aid=15720. [Accessed: 28-Apr-2015].

[73] J. Iserte, F. L. Simonetti, D. J. Zea, E. Teppa, and C. Marino-Buslje, “I-COMS: Interprotein-COrrelated Mutations Server,” *Nucleic Acids Res.*, Jun. 2015.

[74] Z. Li *et al.*, “CorMut: an R/Bioconductor package for computing correlated mutations based on selection pressure.,” *Bioinforma. Oxf. Engl.*, vol. 30, no. 14, pp. 2073–2075, Jul. 2014.

[75] C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, and P. Csermely, “Network analysis of protein dynamics,” *Vienna Spec. Issue Mol. Mach.*, vol. 581, no. 15, pp. 2776–2782, Jun. 2007.

[76] S. Wu *et al.*, “Combined Use of Genome-Wide Association Data and Correlation Networks Unravels Key Regulators of Primary Metabolism in Arabidopsis thaliana,” *PLOS Genet.*, vol. 12, no. 10, p. e1006363, Oct. 2016.

[77] A. Batushansky, D. Toubiana, and A. Fait, “Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism,” *BioMed Res. Int.*, vol. 2016, p. e8313272, Oct. 2016.

[78] D. Toubiana *et al.*, “Correlation-Based Network Analysis of Metabolite and Enzyme Profiles Reveals a Role of Citrate Biosynthesis in Modulating N and C Metabolism in Zea mays,” *Front. Plant Sci.*, vol. 7, p. 1022, 2016.

[79] K. V. Brinda and S. Vishveshwara, “A Network Representation of Protein Structures: Implications for Protein Stability,” *Biophys. J.*, vol. 89, no. 6, pp. 4159–4170.

[80] E. Estrada, “Universality in Protein Residue Networks,” *Biophys. J.*, vol. 98, no. 5, pp. 890–900, Mar. 2010.

[81] D. J. Watts and S. H. Strogatz, “Collective dynamics of /`small-world/’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[82] O. Gaci and S. Balev, “The Small-World Model for Amino Acid Interaction Networks,” in *2009 International Conference on Advanced Information Networking and Applications Workshops*, 2009, pp. 902–907.

[83] O. Gaci, “A Topological Description of Hubs in Amino Acid Interaction Networks,” *Adv. Bioinforma.*, vol. 2010, p. 257512, 2010.

[84] X. Du *et al.*, “Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution,” *Genome Res.*, vol. 18, no. 1, pp. 178–187, Jan. 2008.

[85] “Anaconda Overview,” *Continuum*. [Online]. Available: https://www.continuum.io/anaconda-overview. [Accessed: 21-Feb-2017].

[86] “NumPy — NumPy.” [Online]. Available: http://www.numpy.org/. [Accessed: 21-Feb-2017].

[87] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.

[88] “Neo4j Graph Database: Unlock the Value of Data Relationships,” *Neo4j Graph Database*. .

[89] “PyCharm,” *Wikipedia*. 07-Feb-2017.

[90] “PyCharm :: Features.” [Online]. Available: https://www.jetbrains.com/pycharm/features/. [Accessed: 21-Feb-2017].

[91] “Flask - Full Stack Python.” [Online]. Available: https://www.fullstackpython.com/flask.html. [Accessed: 21-Feb-2017].

[92] “GitHub - pallets/flask: A microframework based on Werkzeug, Jinja2 and good intentions.” [Online]. Available: https://github.com/pallets/flask. [Accessed: 21-Feb-2017].

[93] “GitHub - networkx/networkx: Official NetworkX source code repository.” [Online]. Available: https://github.com/networkx/networkx. [Accessed: 21-Feb-2017].

[94] “Overview — NetworkX.” [Online]. Available: https://networkx.github.io/#. [Accessed: 21-Feb-2017].

[95] M. Bastian, S. Heymann, M. Jacomy, and others, “Gephi: an open source software for exploring and manipulating networks.,” *ICWSM*, vol. 8, pp. 361–362, 2009.

[96] Bokeh Development Team, *Bokeh: Python library for interactive visualization*. 2014.

[97] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[98] “The Py2neo v3 Handbook — The Py2neo v3 Handbook.” [Online]. Available: http://py2neo.org/v3/. [Accessed: 21-Feb-2017].

[99] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, “minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers,” *Bioinformatics*, vol. 29, no. 3, pp. 407–408, Feb. 2013.

[100] “Sigma js.” [Online]. Available: http://sigmajs.org/. [Accessed: 21-Feb-2017].

BIOGRAPHY

Uday Yallapragada received his Bachelor of Engineering degree from GITAM University, Visakhapatnam, India in 1996. He then completed his Masters in Computer Science from Michigan State University, East Lansing in 1998. He has worked as a software developer and software consultant for over 17 years. He is currently a data scientist.