UNIVERSITY OF CAPE COAST

A SYSTEMS LEVEL BASED MODEL FOR IDENTIFYING POTENTIAL TARGETS ASSOCIATED WITH INFLUENZA A INFECTION

LOIS AKU SELASE KEH

2016

UNIVERSITY OF CAPE COAST

A SYSTEMS LEVEL BASED MODEL FOR IDENTIFYING POTENTIAL TAARGETS TARGETS ASSOCIATED WITH INFLUENZA A INFECTION

BY

LOIS AKU SELASE KEH

Thesis submitted to the Department of Mathematics and Statistics of the School of Physical Sciences, College of Agriculture and Natural Sciences, University of Cape Coast, in partial Fulfilment of the requirements for the award of Master of Philosophy degree in Mathematics

JULY 2016

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature:.. Date:...........................

Name: Lois Aku Selase Keh

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature:…………… Date:.........................

Name: Dr. Gaston Mazandu

Co-Supervisor's Signature: ....Date:.........................

Name: Dr. Henry Amankwah

ABSTRACT

Developing therapeutics for infectious diseases requires understanding the main processes driving host and pathogen through which molecular interactions influence cellular functions. The outcome of those infectious diseases, including influenza A (IAV) depends greatly on how the host responds to the virus and how the virus manipulates the host, which is facilitated by protein-protein functional inter-actions and analyzing infection associated genes at the systems level, which may enable us to characterize specific molecular mechanisms which allow the virus of influenza A strains H1N1 and H3N2 to persist and survive inside the host. The system level analysis based on experimental and computational approaches was used to predict human protein-protein functional inter-actions. This human protein-protein functional interaction is a graph consisting of nodes which are proteins, and links joining them. Using this graph, we analyse topological properties of this human protein-protein functional interactions, identify candidate proteins using centrality measures and a map set of IAV infection associated proteins to elucidate genes related to IAV infection and identify essential dense sub-graphs underlying IAV infection outcome. We performed functional closeness and enrichment analyses to identify statistically and biologically significant processes and pathways implicated in IAV infection. These IAV infection associated proteins have shown to be relevant for further research towards new drugs and vaccine development. This study enhances our understanding on the interplay between influenza A and its host and may contribute to the process of designing novel drugs.

KEYWORDS

Drug repositioning

Gene

Influenza A

Key Proteins

Potential Drug Targets

Protein-Protein Interaction

ACKNOWLEDGMENTS

My first thanks go to my supervisors Dr. Gaston Mazandu, thank you for your guidance, constructive criticism, and direction, and to Dr. Henry Amankwah, your honest judgments and directives are very much appreciated. Thank you Rhoda Mahamah for all your help and support, you have been wonderful.

Also, to Mr. Albert Sackitey, thank you for every help you extended to me, I am grateful. My warmest gratitude goes to my parents, pastors, lecturers and friends and everyone that has in one way or the other played a role in my life and brought me this far, God bless you all.

DEDICATION

This work is dedicated to my family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

This thesis presents a system level analysis for the identification of potential targets of Influenza A disease. In this chapter, the background of the study is presented, providing an overview of the whole work. We formulate the problem and discuss approaches that will be used to tackle it, highlighting advantages of these approaches. In summary, this chapter provides a global view of the other chapters.

Background to the Study

Influenza A is a viral disease that can be found in humans and other mammals. According to Smith, (2009), a new influenza A virus which originated from swine surfaced in Mexico and United States in March and early April in the year 2009. In Smith et al. (2009), it is shown that influenza A was derived from several viruses circulating in swine, which lead to the first transmission to humans. The virus had the potential to become first influenza pandemic of the 21st century observed in the year 2009 (Smith, 2009). During the outbreak of the influenza, the first week of surveillance revealed the spread of the virus in over 30 countries through transmission from human to human. This lead the World Health Organization to increase the pandemic alert to level 5 of 6. The different pandemic levels, as described by the World Health Organization (WHO) in the year 2009 are shown in Table 1.

Table 1: Pandemic Levels by WHO

Phases	Description
One	No animal influenza virus circulating among animals has been reported to cause infection in humans.
Two	An animal influenza virus circulating in domesticated or wild animals is known to have caused infection in humans and is therefore considered a specific potential pandemic threat.
Three	An animal or human-animal influenza re-assortant virus has caused sporadic cases or small clusters of disease in people, but has not resulted in human-to-human transmission sufficient to sustain community-level outbreaks.
Four	Human-to-human transmission of an animal or human-animal influenza re-assortant virus able to sustain community-level outbreaks has been verified.
Five	The same identified virus has caused sustained community level outbreaks in two or more countries in one WHO region.
Six	In addition to the criteria defined in Phase 5, the same virus has caused sustained community level outbreaks in at least one other country in another WHO region.
Post Peak	Levels of pandemic influenza in most countries with adequate surveillance have dropped below peak levels.
Possible New Wave	Level of pandemic influenza activity in most countries with adequate surveillance rising again.
Seasonal Influenza	
Post-Pandemic	Levels of influenza activity have returned to the levels seen for seasonal influenza in most countries with adequate surveillance.

The results from the research by Smith, (2009) stated that there was a need for systematic surveillance of the swine influenza and provided evidence that the mixing of new genetic elements in swine has the ability to cause the emergence of viruses with pandemic potential in humans. In medicine, the detection of diseases, the treatment and prevention of many diseases have improved tremendously. This is, in part, due to an improved understanding of biological systems and different factors that trigger progression to diseases.

This has been influenced by advances in high-throughput biological experiments able to generate genome scale datasets of biological cells, including protein sequences, protein-protein interactions, gene expression, regulation and other functional datasets. This has enabled a paradigm shift from single gene analysis to systems level analysis, providing a global view of systems' behavior. This requires the use of systematic mathematical models in order to deal with this large volume of datasets for effective biological knowledge discovery. The system level analysis based on protein-protein interactions has been vital to understanding how proteins function within the cell (Deng and Sheng, 2007). Understanding protein interactions in a given cellular proteome, sometimes known as the *interactome*, is important to the analysis of the cell biochemistry (Deng and Sheng, 2007). A comprehensive collection of information related to human proteins, their features, and their functions is required to ensure information retrieval and possible biological knowledge discovery. For an effective biological knowledge discovery, there is a need to better understand functional activities of proteins in cells and the exact sub-cellular localization of proteins and their tissue-specific distribution. In addition, the knowledge on proteins encoding disease-associated genes play their roles in molecular complexes and biological pathways is very important (Deng and Sheng, 2007). Some facts about influenza A documented by the Center for Disease Control (CDC, 2009) are as follows:

Influenza A is a respiratory disease of pigs caused by type A influenza virus that regularly causes outbreaks of influenza in pigs. Swine influenza viruses may circulate among swine throughout the year, but most outbreaks occur during the late fall and winter months similar to outbreaks in humans.

The classical influenza type A H1N1 virus was first isolated from a pig in 1930. Over the years, different variations of swine flu viruses have emerged. At this time, there are four main Influenza A virus subtypes that have been isolated in pigs: H1N1, H1N2, H3N2, and H3N1. However, most of the recently isolated influenza viruses from pigs have been H1N1 viruses.

Influenza A viruses do not normally infect humans. However, sporadic human infections with swine flu have occurred. Most commonly, these cases occur in persons with direct exposure to pigs, for example, children near pigs at a fair or workers in the swine industry. In the past, CDC received reports of approximately one human swine influenza virus infection every one to two years in the U.S., but from December 2005 through February 2009, 12 cases of human infection with swine influenza was reported. The symptoms of Influenza A flu in people are expected to be similar to the symptoms of regular human seasonal influenza and include fever, lethargy, lack of appetite and coughing. Some people with swine flu also have reported runny nose, sore throat, nausea, vomiting and diarrhoea.

Influenza viruses can be directly transmitted from pigs to humans and from humans to pigs. Human infection with flu viruses from pigs is most likely to occur when people are in close proximity to infected pigs, such as in pig barns and livestock exhibits housing pigs at fairs. Human-to-human transmission of swine flu can also occur. This is thought to occur in the same way as seasonal flu occurs in humans, which is mainly person-to-person transmission through coughing or sneezing of people infected with the influenza virus. Humans may become infected by touching something with flu viruses on it and then touching their mouth or nose. The H1N1 swine flu viruses are antigenically very different from human H1N1 viruses and, therefore, vaccines for human seasonal flu would not provide protection from H1N1 swine flu viruses. In this project, we use systems level computational approaches to identify potential targets of influenza A H1N1 and H3N2. We used a graph-based model to elucidate relationships between different targets identified. Moreover, we performed biological and pathways enrichment analyses. These will produce sub network enriched process and pathways that may play initial role in influenza pathogenesis. In this study, different datasets are derived from different sources to build the protein-protein functional network and perform further analyses. For generating different scores, among the many methods is the application of Information theory, which is a branch of applied mathematics, electrical engineering and computer science and involves the quantification of information. According to Rieke and Warland (1997), information theory was developed by Claude E. Shannon.

In information theory, a candidate measure is *entropy*, which quantifies the uncertainty, which is involved in predicting the value of a random variable. This measure is used at a point in scoring of sequence data. Information theory is based on probability theory and statistics. According to Reza (1961), *entropy* is an important quantity of information and it is common to have a measure between two random variables. A property of entropy is that; it maximizes with a uniform distribution. The entropy H of a random variable Y is associated with measuring intuitively the amount of uncertainty of Y when only the distribution of Y is known Reza (1961). In the same way, entropy is used in this research work to maximize the information content of this work.

Other methods used involve an application of Network theory. In network science and computer science, network theory is the study of graphs as a representation of asymmetric relationships between discrete objects in a general sense. Network theory is also part of Graph theory Newman (2003a). Network theory can be applied in many different areas of study. Some of these areas are statistical physics, computer science, electrical engineering, operations research, gene regulatory networks, and so on. The first true proof in network theory is Euler's solution of the Seven Bridges of Königsberg problem Newman (2003a) which was to devise a walk through the city that by crossing each bridge once, and the starting and ending points of the walk do not need to be the same Newman (2003). There are different types of networks that can be analysed. The social network for instance examines the structure of relationships between social bodies or entities. Persons, groups, organizations, Nation states, websites can be considered as entities in this case. Social network analysis has over the years played a major role in social science. It has been used to analyse several phenomena, including the spread of diseases, the study of markets and many others Wasserman (1994).

In Biological network, the analysis of molecular networks has become central; this is due to the public availability of high throughput biological data, especially protein-protein interaction and other functional datasets. The type of analysis here is almost the same as that of social network analysis, but it focuses on the local patterns in the network. The analysis of biological networks in relation to diseases led to the development of network medicine as another area of application Barabási and Gulbache (2011). Centrality measures which are mostly used in network theory are used in this study to analyze a human-human protein network which is generated.

The knowledge of clustering is also required in this research work. Cluster analysis or clustering consists of grouping objects such that objects in the same group (cluster) are more similar to each other than objects in another group or cluster Bailey (1994). The use of clustering is common in data mining, statistical data analysis, bioinformatics, pattern recognition, image analysis and so on Bailey (1994). Clustering has no specific algorithm that can be used. Clustering can be done by different algorithms both in notion and in how to efficiently find clusters. Some of the algorithms are, the agglomerative algorithm that merge similar nodes recursively, the and divisive algorithm which detects inter-community links and remove them from the network. These methods do not produce a unique partitioning of the data set; they however produce a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge, hence they are too slow for large datasets. There is however another algorithm introduced by Blondel and Guillaume (2008) which is what we use in this work, reasons being that it is fast and can produce quick results for a large network unlike the agglomerative, and divisive algorithms.

Statement of the Problem

The World Health Organization (WHO) in the year 2009 announced that the influenza A pandemic was over and the disease had now returned to a seasonal flu state. Research by Smith et al. (2009) indicated that influenza A virus had the potential to cause the emergence of viruses with pandemic potential in humans. Furthermore, CDC stated that when influenza viruses from different species infect pigs, the viruses can re-assort, that is swap genes and new viruses that are a mix of swine, human and/or avian influenza viruses can emerge. Over the years, different variations of swine flu viruses have emerged. At this time, there are four main Influenza A virus subtypes that have been isolated in pigs:H1N1, H1N2, H3N2, and H3N1. However, most of the recently isolated influenza viruses from pigs have been H1N1 viruses. There also exists a seasonal outbreak of influenza A, especially during winter in several countries worldwide. The major problem is the fact that the virus has the ability to reassort and become very risky with no available treatment. It is in view of this that Smith et al. (2009) recommended a systematic surveillance and also the case of the seasonal outbreak. This elicits the need for optimally predicting and identifying potential targets of influenza A disease in order to design effective drugs for treating the disease.

Objectives of the Study

The objectives of this study are to identify potential targets of influenza A using a system level based model. This is achieved as described as follows:

1. Generate a human protein-protein functional network.
2. Identify the protein targets of influenza A.
3. Filter differentially abundant proteins and perform an enrichment analysis through Gene Ontology (GO) annotations and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG).
4. Predict potential drugs.

Significance of the Study

The detection of diseases from the early stages and understanding how they operate within the human are essential in setting optimal and therapeutic strategies. Most ailments caused by viruses and bacteria resurface from time to time. This is also the case for influenza A, which has returned to a seasonal state and some other infections like Ebola, which seem to have disappeared. The identification of influenza A virus (IAV) potential targets using a systems level analysis and predicting potential drugs that can match these targets are being used in this study. It is worth mentioning that, this approach can be generalised to other infections even though it is used for IAV infection.

Delimitation

This work is limited to proteins sequences; it does not take into account protein-protein interactions between host and human outside the host. It considers the interactions that take place within the host only.

Limitation

The prediction of drugs requires the use of semantic similarity measures, the prediction also makes use of a threshold with a tuning parameter, which is not fixed, and hence, the tuning parameter has to be adjusted until the threshold produces desired results.

Definition of Terms

Some terms that will be seen in this research work are defined in this section.

Definition one: Target or target protein also known as a candidate protein in this research work. A candidate protein is a protein that plays a major role in the network and helps to identify the potential targets of the influenza A disease.

Definition two: Surveillance is the ongoing systematic collection and analysis of data about an infectious disease that can lead to action being taken to control or prevent the disease.

Definition three: A gene is the basic physical and functional unit of heredity. Genes are made up of DNA and act as instructions to make molecules called proteins. The genes in humans vary in size, that is, from a few hundred DNA bases to more than two million bases.

Definition four: Proteins are biomolecules that consist of one or more chains of amino acid residues. Proteins differ from one another in their sequence of amino acids.

Definition five: Sequencing means determining the primary structure of a biopolymer which is not branched in genetics. There are different types of sequencing such as DNA sequencing, RNA sequencing, and protein sequencing.

Definition six: Protein Sequencing is a method used to determine the amino acid sequence of a protein, as well as which conformation the protein adopts and how complex it is with any non-peptide molecules.

Definition seven: Genome is a complete set of DNA of an organism, and this includes all of its genes.

Definition eight: Gene Ontology(GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. There is more information on Gene Ontology in Chapter Three.

Definition nine: Enrichment analysis is a method of identifying classes of genes or proteins that are over-represented in a large set of genes or proteins.

The Expected Outcomes

At the end of this research, we expect the following outcomes:

1. Candidate proteins associated with Influenza A.
2. Drugs showing higher functional similarity to candidate proteins.

Organization of Study

The rest of this study is organized as follows:

In Chapter Two, we present a review of previous work done on influenza A disease, concentrating on methods and findings of each paper that will be reviewed. Chapter Three outlines in detail the materials used in this research, explaining the functional datasets which were used and an insight on how they were extracted for use, Chapter Three also presents the integrative model employed in this research work. Chapter Four presents all results and findings in this research work, and in Chapter Five, we present the summary, and conclusion of the whole work, and recommendations.

Summary

In this chapter we see a background of this study, and the problem statement where we establish the need to conduct this research and how the findings can be beneficial to society. Also, we mentioned that the methods presented in this research work could be directed to other diseases. The objectives were to find potential targets of influenza A also to predict drugs which can be repositioned for treating influenza A.

CHAPTER TWO

LITERATURE REVIEW

Introduction

Influenza A is a very contagious respiratory disease of pigs and is caused by the type A influenza virus, which regularly causes outbreaks of influenza in pigs. Influenza A virus is an *orthomyxovirus* and contains the glycoproteins *haemagglutinin* and *neuraminidase* Lim and Mahmood (2011). Influenza A can be described as H1N1, H1N2 depending on the type of H or N antigens that is expressed metabolically. Haemagglutinin causes red blood cells to clump together and binds the virus to the infected cell. Neuraminidase (N), on the other hand, is a type of glycoside hydrolase enzyme which helps the particles of the virus to move through the infected cell. It also assists in the budding from the host cells Lim & Mahmood (2011). Influenza A viruses do not normally infect humans. However, human infections have occurred and the 2009 pandemic is a clear risk indicator of the exposed people getting infected. There have also been cases of human to human spread of swine flu. The flu is spread in the same way as seasonal flu in humans. This chapter presents a review on both influenza A and its subtypes

Reviews on Influenza A

In this section, we review some relevant literature on influenza A in relation to this study. Van Kerkhove and Vandemaele (2011) gathered information to characterize the severity of H1N1 pandemic (H1N1pdm) infection and to assist policy makers to identify risk groups for target control measures. They compared data that was primarily obtained from surveillance programs of the Ministries of Health or National Public Health Institutes of 19 countries from 1st April 2009 to 1st January 2010 which is a 10-month period. These countries were requested to provide risk factor data on laboratory-confirmed cases according to a standardized format they used for analyses. The data used was collected during the period of routine surveillance and this varied from country to country. The data was reported to the team with no identification. It was also reported as aggregate data to avoid issues related to ethics approval. Furthermore, since many countries did not agree to have their country-specific data published together with other countries, they had to approach the publishing of data considering a wide range of countries and showed the variability observed so that results from specific studies can be compared with international results reported in their work. They grouped the potential risk factors into four categories: age, chronic medical illnesses, pregnancy which was recorded by trimester, and “other”. The “other” comprises of conditions that were not considered as risk factors for severe influenza outcomes. Some of these “other” factors are: obesity, membership in a vulnerable social or ethnic group, and TB Van Kerkhove et al. (2011). Also, this risk factor information was collected separately for the three levels of severity of the illness in laboratory-confirmed patients: hospitalizations, admissions to Intensive Care Unit (ICU), and fatalities by country. Thus they calculated the percentage of patients who were hospitalized, those who were admitted to ICU and died using the total number of cases reported in each severity category with the exception of pregnancy Van Kerkhove et al. (2011).

In order to evaluate risk associated with pregnancy, the ratio of pregnant women to all women of child bearing age (15-49 years) in each level of severity was used to describe the difference between the levels. They evaluated the overall median and interquartile ranges (IQRs) for each risk factor using all the data that was available. According to Van Kerkhove et al. (2011) countries also provided a baseline comparison data or prevalence of the risk factor in the general population where it was available. The data on age was provided by the following groups: $< 5{,}5-14{,}15-24, 25-49, 50-64$ and ≥ 65. The countries that were used in the analysis include: Argentina, Australia, Canada, Chile, China, France, Germany, Hong Kong SAR, Japan, Madagascar, Mexico, The Netherlands, New Zealand, Singapore, South Africa, Spain, Thailand, The United States and the United Kingdom.

They also used I^2 statistic to quantify the percentage of variation across countries that is due to true underlying heterogeneity in the odd ratios. I^2statistic is an alternative approach for quantifying the effect of heterogeneity. It also provides a measure of the inconsistency in the results of studies', and is used to describe the percentage of total variation across studies, which is a result of heterogeneity rather than chance Higgins et al. (2003). Results indicated that, the median age of patients correlated to the disease's severity. The hospitalization risk per person was among patients under 5 years and between 5 to 14 years. The highest risk of death per person was between the age groups of 50-64 years and above. Van Kerkhove et al. (2011) observed that morbid obesity might be as risk factor of ICU admission and fatal outcome. They concluded that the risk factors for H1N1pdm are similar to those of seasonal influenza with obvious differences, such as younger age groups and obesity. They stressed on the need to identify and protect groups at highest risk of severe outcomes. These results suggested that there is a need for further studies toward implementation of optimal measures and strategies to counter influenza A in case of a new outbreak.

Meijer and Lackenby and Hungnes (2009) reported an oseltamivir-resistant influenza A virus which were with H275Y mutation in the neuraminidase emerged independently of drug use. Country by country, the proportion of oseltamivir resistant viruses (ORVs) ranged from zero percent (0%) to sixty eight percent (68%) with the highest proportion in Norway. It was discovered in January 2008 by Meijer et al. (2009) that there was an unexpected high rate level and spread of oseltamivir resistant influenza A virus in Europe. This was long before the outbreak in 2009. The methods employed by \cite{resistant} included the use the European Influenza Surveillance Scheme (EISS) which monitored influenza activity from October 1-7 of the year 2007 to May 5-11 of the year 2008. The EISS covered all 27 European Union Countries, and Croatia, Norway, Serbia, Switzerland, Turkey, and Ukraine. In each country, there were one or several networks of sentinel general practitioners (GPs) reported rates of consultation for influenza-like illness (ILI) or acute respiratory infection (ARI). ARI includes ILI and all other infections. No consultation data were available for Croatia, Finland, Turkey and Ukraine, all this was for the clinical influenza activity. There was also a virologic analysis, which was accomplished with the help of sentinel GPs who forwarded nasal and pharyngeal specimens from a subset of their patients to the National Influenza Center (NIC) for virus detection. Specimens and influenza viruses collected from other sources (hospitals, non-sentinel) were also examined by the NIC. Turkey and Cyprus had no virus detection data. An antiviral drug susceptibility data was also generated via the European Surveillance Network for Vigilance against Viral Resistance (VIRGIL) project at the UK Health Protection Agency laboratory in London or by individual NICs. A genetic analysis was performed, using cycle-sequencing on clinical specimens. Data on Antiviral susceptibility was not available for Cyprus, Lithuania and Malta. The data collected was analyzed as follows: Clinical, virologic surveillance, and antiviral susceptibility data for England, Northern Ireland, Scotland and Wales were totaled in order to obtain estimates for the United Kingdom Meijer et al. (2009).

The results of Meijer et al. (2009) captured the 2007 to 2008 influenza season in Europe, which was initially dominated by influenza viruses A ($n = 10{,}720$; 60% of all influenza virus detections). Influenza viruses B ($n = 7{,}150$; 40% of all influenza virus detections) became dominant in week 8 according to results obtained. Of the 5,984 (56%) influenza viruses A subtyped, 5,748 (96%) were found to be H1, and 236 (4%) were found to be H3. In total, influenza virus detections were at a peak in week 6.It took 4 weeks for influenza viruses A (H1N1), 8 weeks for influenza viruses B. Out of the 2,136 influenza viruses A (H1N1) which were characterized antigenically, 97% were reported to be “closely related to the vaccine strain A/SolomonIslands/3/2006”, even though half of the number of these viruses were reported to be more “closely related to A/Brisbane/59/2007", the vaccine strain recommended for the 2008–2009 season Meijer et al. (2009).

Presanis et al. (2009) studied on the severity of pandemic H1N1 influenza in the United States from April to July, 2009. Measuring the severity of pandemics such as H1N1 accurately is very important. The pandemic in 2009 resulted in over 209,000 laboratory confirmed cases of H1N1 and 3205 plus deaths worldwide. Many national and international authorities acknowledged that the numbers they published were substantially underestimated and were a reflection of an inability to identify, test, confirm and report many cases. There are many ways of measuring the severity of infection and one of the simplest is the case-fatality ratio (CFR), which is the probability that an infection causes death Presanis et al. (2009). There are other measures of severity that are pertinent to the burden a pandemic exerts on the health care system. These are the case-hospitalization and the case-intensive care ratios (CHR and CIR) respectively, which computes the probabilities that an infection can lead to hospitalization or admission to the intensive care unit Presanis et al. (2009).

When it comes to estimating severity of pH1N1 infection, it includes the problem of estimating the number of infected individuals within a given population and the time period to develop symptoms are medically attended, hospitalized, admitted to ICU, and die because they have been infected. No large legal power in the world was able to keep an accurate number or maintain the total number of pH1N1 cases as soon as the epidemic grew above hundreds of cases Presanis et al. (2009). Hence, the main goal of Presanis et al. (2009) was to estimate the probability of hospitalisation, ICU admission, and death by age group, and overall for each symptomatic pH1N1 case. The challenge was that, for any large population, to have a significant number of patients with severe outcomes, they had no reliable measure of the number of symptomatic pH1N1 cases, therefore, the problem was tackled in two ways. The first approach was to view the severity of infection as a ``pyramid", where each successive level represented a greater severity, in order to estimate the ratio of the top level to the base which is also the symptomatic cases, the ratios of each successive level to the one below it was estimated. See left side of Presanis et al. (2009)

Thus, the sCFR was split, that is, the probability of death per symptomatic case, into components for available data, the probability of a case coming to medical attention given symptomatic infection from the CDC survey data; the probability of being hospitalized given medical attention from the Milwaukee data; and the probability of dying given hospitalization from the New York data, which included a correction for those who died of pH1N1 but were not hospitalized. The second approach used the self-reported incidence of ILI from a telephone survey in New York City as the estimate of total symptomatic pH1N1 disease, and the total number of confirmed deaths in New York City as the estimate of the deaths, that was imperfect ascertainment was accounted for, in this case because of the possibility of imperfect viral testing sensitivity. All the estimates were combined within a Bayesian evidence synthesis framework. The framework allowed the estimation of probabilities for the quantities of interest such as the sCFR, sCIR, and sCHR, and associated uncertainty which is expressed as credible intervals [CIs].

Credible intervals, gives an appropriate picture of the combined uncertainties which are associated with each of the inputs to the estimate. This is mainly the true counts of cases at each level of severity, and this is after imperfect detection and the uncertainties due to sampling error (chance) has been accounted for Presanis et al. (2009).

In approach one of Presanis et al. (2009), the New York and Milwaukee data is combined for the unobserved level of severity (symptomatic cases). The main analysis of the first approach was performed using prior information to inform the detection of probabilities. In addition, a naive analysis was performed, and the detection probabilities, *d*, were set to be equal to 1 at all levels of severity. The prior distributions for the number of symptomatic cases in New York in general and according to age were taken as a uniform range between zero and the proportion reporting ILI in the telephone survey which also had an upper bound of that distribution itself having a prior distribution which reflects the confidence bounds of the survey results Presanis et al. (2009).

In approach two of Presanis et al. (2009), the New York case data and telephone survey data was used. The assumption that self-reported ILI cases represented symptomatic pH1N1 infection was made. The mean and 95% confidence intervals from the survey was used to define a prior distribution on the number of symptomatic cases all together and according to age group. Observed hospitalizations, ICU/ventilator use, and fatalities were used along with prior distributions on detecting probabilities. This was to inform estimates of true numbers of hospitalizations, ICU/ventilator use, and fatalities, which were then used to estimate sCHR, sCIR, and sCFR Presanis et al. (2009).

From approach, one, two alternatives were considered to estimate the ratios of interest from the combined New York and Milwaukee data, making use of self-reported rates of medical attention seeking to establish the denominator. Firstly, the naive estimate of the ratios of deaths to hospitalizations, ignoring differences in detection across levels of severity were obtained; secondly, an estimate that included evidence and expert opinion on the detection probabilities at each level of severity was obtained. The naive estimate suggested a median ratio of deaths to hospitalizations of 4.3%, ICU admissions to hospitalizations was 25%, and of hospitalizations to medically attended cases was 3.1%. Also, the ratio of deaths outside of hospitals to medically attended cases was estimated to be 0.03% Presanis et al. (2009). From approach two, the overall estimates are: $\mathrm{sCFR} = 0.007\%, \mathrm{sCIR} = 0.028\%$ and $\mathrm{sCHR} = 0.16\%.$ Comparing approach one to approach two, the estimates are almost an order of magnitude smaller, and the age distribution is different. The respective risks for each severity in the 18 to 64-year-old group compared to the 5 to 17-year-old group are 7 for fatalities, 1.5 for ICU admissions, and 1.4 for hospitalizations. The CFR is highest in the 18–64-year-old group with a posterior probability of 52% Presanis et al. (2009). Presanis et al. (2009) estimated data from two cities on tiered levels of severity and self-reported rates of seeking medical attention. An approximate value of 1.44% of symptomatic pH1N1 patients during the spring in the US were hospitalized, 0.239% of them required intensive care or mechanical ventilation; and 0.048% died. Within the assumptions Presanis et al. (2009) made in the model presented, the estimates were uncertain up to a factor of about 2 in both direction, as reflected in the 95% credible intervals associated with the estimates Presanis et al. (2009).

Presanis et al. (2009) concluded that, the estimated severity is an indicator of a satisfactory expectation for the autumn–winter pandemic wave in the US was a death toll less than or equal to that which was typical for seasonal influenza, even-though there was a possibility with a considerable large amount of deaths in younger persons. It was then suggested that continued close monitoring of the severity of pandemic H1N1 influenza A was needed in order to assess how patterns of hospitalization, intensive care utilization, and fatality were varying in space and time and across age groups. An increases in the severity was likely to reflect changes in the host population; for example, infection of persons with conditions that predispose them to severe outcomes, or changes in the age distribution of cases; an example was a shift toward adults, in whom infection will be more severe. Changes in severity could also reflect changes in the virus or duration in the access and quality of care available to infected persons Presanis et al. (2009).

In the year 2000, as the awareness of the threat of a new influenza A pandemic increased, Fouchier et al. (2000) set out to propose a method of detecting influenza A virus from different species. Virus isolation was not an option because of its underlying limitation in sensitivity, and lack of host cells that are generally permissive to all influenza A viruses Fouchier et al. (2000). Migratory birds and water fowls are considered to be the natural storehouse of influenza A virus. Influenza A viruses which represent 15 *hemagglutinin* (HA) and nine *neuraminidase* (NA) subtypes were detected in wild birds and poultry worldwide. The general human populations are said to be serologically naive in relation to most avian HA and NA antigens and hence avian originated influenza A viruses posed as a threat, this formed the basis of the new influenza A pandemic in humans Fouchier et al. (2000). Until the zoonotic events in Hong Kong and China caused by H5N1 and H9N2 influenza viruses which suggested a possible direct transmission of avian related influenza viruses, it was always thought that avian influenza viruses could be transmitted to humans only via co-infection and genetic reassortment of avian and swine or human influenza viruses in pigs Fouchier et al. (2000). Although there were procedures such as vitro virus isolation and immunofluorescence (IF) for detecting human influenza A virus, they were less effective in detecting influenza A viruses of avian and swine origin. This was because, the phenotypic and genetic heterogeneities of the newer viruses may result in false-negative diagnosis of influenza A virus infection by in vitro cell culture Fouchier et al. (2000). The aim of Fouchier et al. (2000) was to setup a rapid and sensitive Polymerase Chain Reaction (PCR) method for the screening of clinical specimen to if there are phenotypically and genotypically diverse influenza A virus. Hence, a primer set for PCR based detection of influenza A viruses was designed. This was validated using strains which represented all known HA and NA subtypes which were extracted from different host species at different geographic locations. In ensuring that the PCR based method had the capacity to produce the desired results, PCR based screening samples from human and avian origin was compared with the classical isolation of influenza A virus in the mammal cell culture. The conclusion was that the PCR method proposed by Fouchier et al. (2000) in the detection of influenza A virus was fast, sensitive, and specific and was suitable for all genetic variants of influenza A viruses known as at that time Fouchier et al. (2000).

The procedures that lead to this conclusion is as follows; PCR primers were designed based on sequence data obtained from the influenza sequence database at Los Almos National Laboratory, Los Almos. A Bio edit software package was used to identify the conserved sequence in the influenza virus gene segments by creating entropy plots. Partial sequences which represents gene segments 5, 7 and 8 encoding nucleoprotein, matrix and non-constructural proteins respectively were analysed. This is because HA and NA genes are genetically diverse and the sequence information on the PA, PB1 and PB2 polymerase genes was limited, that is to say, less than 100 sequence entries were available from the database, partial sequences inclusive Fouchier et al. (2000). Entropy was the measure used to express the degree of heterogeneity. Cloacal swabs were collected from ducks (gadwall, and mallard), cloacal swabs and droppings were collected from geese (white fronted, barnacle and brent). Duck samples were collected at Lekkerkerk, and goose samples were collected in Gronigen and Eemdijk, all in the Netherlands. Cotton swabs were used for the sampling process of the birds. Finally, throat swab specimens were collected from humans. RNA was then isolated using a high pure RNA isolation kit. The PCR amplification was then applied on the dataset.

Throat swab samples were sent to the virus diagnostic laboratory at Erasmus University Medical Center are routinely tested for the presence of influenza A virus. For a selection of Influenza A virus-positive, throat swab samples which were obtained in the 1994-1995 influenza season. Influenza A virus titers were determined by end point dilution. From a total of 26 negative control samples (13 were influenza B virus positive and 13 were influenza A and B virus negative in mammalian cell cultures), 24 were found to be negative upon PCR and dot blot analyses. Two of the swabs were negative for influenza A virus in mammalian cell culture and by IF but yielded very weak signals after PCR and dot blot hybridization. The weak dot blot signals were likely due to background hybridization or the presence of very small amounts of influenza A virus RNA in the throat swabs Fouchier et al. (2000). The suitability of the PCR for avian influenza A virus screening of cloacal swab and dropping samples from ducks, geese, and shorebirds was also tested. The PCR screening appeared to be up to 100-fold more sensitive than virus isolation and hence reduced cost and workload. Out of the 235 pools of samples which represented 1,175 individual specimens, RNA isolation, PCR, and Southern or dot blot hybridization revealed that influenza A virus was present in 19 of them. For each of the individual samples present in these 19 pools, RNA was isolated from them and it was revealed that all except 1 pool contained a single positive bird sample, however, the one exception contained two positive samples Fouchier et al. (2000). Hence, Fouchier et al. (2000) concluded that the PCR based method was better at detecting influenza A in humans and animals. Today, there are modern and highly effective methods of detecting viruses and other infections. One of such methods is the system level analysis which is applied in this research work.

Since its detection in April 2009, an A (H1N1) virus which contains a unique combination of gene segments from both North American and Eurasian swine viruses have continued to circulate in humans. The absence of similarity between the 2009 A(H1N1) virus and its nearest relatives is an indication that its gene segments had been circulating without being detected for a very long time. The virus has low genetic diversity and hence, it suggests that the virus was introduced to humans in a single event or multiple events of similar viruses Garten et al. (2009) Influenza pandemics according to Garten et al. (2009), there is little or no existing immunity against an influenza virus with hemagglutinin (HA) emerges in a population and is transmitted from human to human effectively Garten et al. (2009). The genomes of the past pandemic influenza viruses before the 2009 pandemic virus, that is, 1918 H1N1, 1957 H2N2, and 1968 H3N2 all originated wholly or partially from non-human reservoirs. The HA genes of all the pandemic viruses originated ideally from avian influenza viruses Garten et al. (2009). The first isolation of A(H1N1) influenza viruses were from swine. They were shown to be highly antigenetically similar to the reconstructed 1918 A (H1N1) virus and were likely to share a common ancestor. From the year 1930 to the late 1990s, the “classical swine influenza” viruses which circulated in swine and stayed stable antigenetically Garten et al. (2009). Swine became a reservoir of H1N1 with a potential of causing major respiratory outbreaks or possible pandemics, this is because the relative anigentic stasis of classical H1N1 viruses in swine until 1998 when the substantial drift of H1 in humans was observed and created a gap between classical swine H1 and human seasonal H1 viruses. The 2009 A(H1N1) virus is a combination of gene segments that were not reported previously in swine or human influenza virus in the United States or elsewhere Garten et al. (2009).

In conclusion, the circulation of an influenza A(H1N1) swine origin virus in humans with an antigenically and genetically divergent HA and a previously unrecognized genetic composition is of concern to public health officials around the world. The virus appeared to be readily transmissible between humans was a cause for alarm. The evolutionary distances between the gene segments of the A (H1N1) virus and its closest relatives show a lack of surveillance in swine populations that are likely to harbor influenza viruses with pandemic potential. Finally, Garten et al. (2009) suggested that the Worldwide monitoring of the antigenic and genetic properties of the 2009 A(H1N1) viruses should continue for, among other reasons, detecting any changes and thus any necessity for selecting further vaccine candidates or changes in antiviral recommendations.

According to Davis et al. (2014), Influenza A viruses are infectious agents spread through contact or aerosol droplets that result in a seasonal respiratory illness which can potentially lead to death. This has already been stated in previous reviews. The natural reservoir for influenza A viruses are aquatic birds, however, many animals are susceptible to infection, including swine and humans. Even though humans are not readily infected with avian influenza viruses, there are rare cases where a direct avian to human transmission has occurred. Swine on the other hand are readily susceptible to avian, swine, and human influenza subtypes. They also provide a vessel for genome reassortment among different subtypes of the virus. The influenza A virus genome is made up of eight negative sense RNA segments (vRNA) Davis et al. (2014). The reassortment of genome segments between different influenza A subtypes have the tendency to yield new influenza A subtypes that have potential to cause a human influenza pandemic. Even though the 1918 pandemic virus was found to be of wholly avian origin, the 1957 and 1968 pandemics contained segments of avian and human origin which makes it unclear if swine or human were the vessel of reassortment. Swine are more readily infected with avian influenzas, and this provides more opportunity for reassortment of the virus. Humans can also be infected with avian influenza, although the occurrence is rare, reassortment within a human host is still possible Davis et al. (2014). The aim of the review by Davis et al. (2014) was to highlight the current state of antiviral resistance in circulating and highly pathogenic influenza A viruses and to explore the potential antiviral targets within the proteins of the influenza A virus ribonucleoprotein (vRNP) complex, and to draw attention to the viral protein activities and interactions that play an indispensable role in the influenza life cycle. An investigation of small molecule inhibition, accelerated by the use of crystal structures of vRNP proteins, was also conducted and this provided important information about viral protein domains and interactions, and also revealed many potential antiviral drugs.

Prevention of influenza infection is typically by annual vaccination. However, vaccines tend to be useless after infection or against emerging subtypes of influenza which were not targeted during vaccine production. This was witnessed with the novel H1N1 pandemic in 2009. Therefore, antivirals that target specific proteins to inhibit virus replication are very necessary to cater for the spread of an emerging pandemic Davis et al. (2014). Due to the many viral protein interactions required in various stages of the influenza life cycle, there are numerous potential target sites for antiviral treatments. Current antivirals target the activities of M2 and NA, but resistance is emerging. This review catalogs the current state of influenza antiviral resistance and describes potential new molecules, which target proteins within the viral ribonucleoprotein (vRNP), the complex comprises of the viral RNA genome, the RNA-dependent RNA polymerase (RdRP), and nucleoprotein (NP). Two viral segments code for the surface proteins HA and NA for which influenza subtypes are named Davis et al. (2014). ``The current antivirals approved by the FDA are, in order of their release, Symmetra (amantadine), Flumadine (rimantadine), Relenza (zanamivir), and Tamiflu (oseltamivir). Amantadine and rimantadine are adamantane derivatives that target and inhibit the M2 ion channel. The M2 ion channel is an integral membrane protein responsible for release of vRNPs during infection. By binding to the M2 ion channel, amantadine and rimantadine inhibit vRNP release and thus viral replication" Davis et al. (2014).

The antivirals zanamivir and oseltamivir inhibit NA or neuraminidase of which their activity is required to release new virions from infected cells. These drugs inhibit the release of the virus from infected host cells by binding them to the active site of the NA protein Davis et al. (2014). The resistance against both classes of influenza antiviral treatments have been documented. Resistance to M2 ion channel inhibitors also occur through a triple amino acid deletion or single amino acid substitutions in the transmembrane region. One hundred percent (100%) of H3N2 influenza A viruses circulating in 2009–2010 and 99.8% of 2009 pandemic H1N1 were resistant to adamantanes Davis et al. (2014). Resistance to NA inhibitors however is less common as this class of inhibitors were developed later. An example is the 98.9% of tested 2009 H1N1 viruses which remained susceptible to oseltamivir, and 100% of 2009 H1N1 viruses tested also remained susceptible to zanamivir, 100% of influenza A H3N2 which were tested remained susceptible to both oseltamivir and zanamivir for the 2012–2013 season Davis et al. (2014).

A multiple single amino acid changes in NA can alter the susceptibility to the approved neuraminidase inhibitors. Resistance normally evolves during treatment, through single amino acid substitutions which include changes to amino acids such as E119V, I223R, and H275Y. Pandemic H1N1 2009 isolates with a substitution at I223R were resistant to neuraminidase inhibitors in addition to the M2 ion channel inhibitors Davis et al. (2014). Hence, while NA inhibitors are currently still viable to combat most emerging influenza threats, it is only a matter of time before resistance is established as with the adamantanes, and this will render both old and current antiviral therapies ineffective against an emerging influenza threat Davis et al. (2014). What is more disturbing is the resistance reported among Highly Pathogenic Avian Influenza (HPAI) subtypes that could spur the next pandemic Davis et al. (2014).

Davis et al. (2014) establishes that avian subtypes such as H5N1, H7N9, and H7N7 have all resulted in human infection. The H5N1 infections also resulted in high morbidity at approximately 60%, while H7N9 and H7N7 have seen more inconsistency in outcome of human infection Davis et al. (2014). Fortunately, none of the subtypes have acquired the ability to transmit readily from human to human, unfortunately, the strains already have antiviral resistant isolates which have been reported. An example is the fact that all H7N9 isolates tested were resistant to adamantanes through the S31N substitution in the M2 protein, and some H7N9 exhibited high resistance to oseltamivir, mid-resistance to peramivir, and low-resistance to zanamivir through the NA R292K substitution Davis et al. (2014).

The critical roles of the influenza vRNP for viral RNA synthesis makes the activities of the vRNP, such as cap snatching and RNA polymerization, excellent antiviral targets. A recent discovery of nucleotide analog preferably utilized by viral RNA dependent RNA polymerases including influenza vRNP, was under study as a potential antiviral therapy by targeting the activity of viral RNA dependent RNA polymerases. In addition, the multiple essential interactions of the vRNP, such as with each other to form an RNA dependent RNA polymerase heterodimer, with host capped mRNAs to obtain primers for viral transcription, and with NP to control and improve RNA replication, together with high conservation of the domains among influenza subtypes, and make the proteins of the vRNP excellent targets for small molecule inhibitors with broad efficiency against multiple influenza A subtypes Davis et al. (2014).

Davis et al. (2014) concluded that influenza A virus continues to remain a human menace, in terms of both human health and global economic costs. The wide host range of influenza A virus, together with an inability to proofread activity within the viral RNA dependent RNA polymerases, and a segmented RNA genome allows for segment reassortment and provides influenza A virus with the ability to rapidly evolve. Even though yearly vaccination protects against the strains and subtypes which have been predicted to be circulating and represented during vaccine production, the vaccination will not protect against an unseen and emerging subtype of the virus. Antivirals happen to be the first line of defense for any emerging pandemic and resistance to current antivirals is already circulating within influenza A virus, quickening the resolve to identify new antiviral therapies. Davis et al. (2014) added that viral ribonucleoprotein (vRNP) is essential for viral replication, making it an ideal target for antivirals Davis et al. (2014).

Davis et al. (2014) continued to state in conclusion that the essential activities of the vRNP include cap-snatching activity and required for viral mRNA transcription and RNA polymerase activity required for viral mRNA transcription and RNA replication, and the interactions required to form functional vRNPs with these essential activities comprised the most highly conserved protein domains within influenza A subtypes. The vRNP provided multiple viral protein targets to reduce the selection pressures and emergence of resistant strains. However, mutations which confer antiviral resistance were tolerated, and history dictates the mutations will be selected by use of the antiviral and propagate in circulating influenza A viruses. This implied the search for new influenza A antiviral inhibitors should be ongoing until a universal vaccine is achieved Davis et al. (2014).

Hindiyeh et al. (2010) reported on the development and validation of high-throughput real-time reverse transcriptase PCR assays for the detection of the H275Y substitution in the neuraminidase 1 gene that can be accomplished in 3 to 4 hours. According to Hindiyeh et al. (2010), the World Health Organization (WHO) has a continuous surveillance for oseltamivir-resistant viruses, due to the number of documented sporadic resistant cases increasing, and reaching nearly 100 cases worldwide by 15 December 2009. The U.S. Centers for Disease Control and Prevention (CDC) also reported the first human-to-human transmission of oseltamivir-resistant pandemic (H1N1) 2009 virus in two summer campers who were receiving oseltamivir prophylaxis. In addition, the WHO announced the outbreaks of oseltamivir-resistant pandemic (H1N1) 2009 virus in two immunocompromised groups, one in North Carolina and the other in Wales, United Kingdom. In both outbreaks, human-to-human oseltamivir-resistant virus transmission was suspected Hindiyeh et al. (2010).

There are at least two mechanisms which contribute to neuraminidase resistance in the seasonal influenza viruses H1N1 and avian influenza H5N1. One of these mechanisms involves reduction of the binding efficiency of virus hemagglutinin to its receptor whilst the other is associated with amino acid substitutions in and around the NA active site, of which the substitution at position 275 (histidine to tyrosine [H275Y]) is very common Hindiyeh et al. (2010). The sequence analysis of the hemagglutinin gene is however not a reliable indicator of neuraminidase drug resistance phenotype, but histidine-to-tyrosine (H275Y) substitution in the active site of the NA-1 gene gives an indication of reduced binding affinity of the neuraminidase inhibitor oseltamivir. Reports that characterize the current oseltamivir-resistant pandemic (H1N1) 2009 virus confirmed the presence of the H275Y mutation Hindiyeh et al. (2010).

The phenotypic analysis of oseltamivir-resistant influenza A virus happens to be widely accepted as the “gold standard” methodology for detecting influenza virus drug resistance. A genotypic analysis has been widely used to detect a point mutation (cytosine to thymine) at position 823 of the NA-1 gene that can result in a histidine-to-tyrosine substitution. The genotypic assays included sequencing a part of the neuraminidase gene by using the Sanger dideoxy sequencing method or by pyrosequencing Hindiyeh et al. (2010). The assays were labor-intensive, and had a long turnaround time ranging from 24 to 72 hours. This required specialized equipment and human effort. Also, sequencing and pyrosequencing assays have reduced sensitivities for detecting low concentrations (<15%) of quasispecies present in a patient's sample Hindiyeh et al. (2010). This called for high-throughput assays with short turnaround times in order to expedite oseltamivir drug resistance detection.

Hindiyeh et al. (2010) then validated two real-time reverse transcriptase PCR (qRT-PCR) assays by utilizing TaqMan chemistry for the detection of the point mutation (cytosine to thymine) at position 823 of the NA-1 gene of pandemic (H1N1) 2009 virus as part of their study. One set of modified primers and two probes previously reported by Chutinimitkul et al. (2007) were utilized to validate both assays. According to the team, Chutinimitkul et al. (2007) initially validated the primers and probes for the detection of oseltamivir resistance in H5N1 isolates. The primers were then used in the present study and were modified to increase the assay's sensitivity and specificity to detect pandemic (H1N1) 2009 virus, while the two minor groove binding (MGB) probes were synthesized per the description of Chutinimitkul et al. (2007). The point of mutation in the probe was positioned 7 bases from the 3' end in order to minimize the cross-reactivity with similar sequences Hindiyeh et al. (2010). Hindiyeh et al. (2010) then investigated specificities of the two assays by analyzing one clinical isolate of each of the common human respiratory viruses that were strongly positive by real-time PCR analysis. These included seasonal influenza A virus (H1N1 and H3N2), influenza B virus, human metapneumovirus, respiratory syncytial virus (RSV) types A and B, and adenovirus type 2 Hindiyeh et al. (2010). The two assays were specific, because no positive signals were noted in the analysis of these viruses Hindiyeh et al. (2010). The validation of the two assays, H275 (sensitive) and H275Y (resistant), was then performed on 31 patient samples collected between 15th June, 2009 and 1st December, 2009 from 18 patients with a high risk of developing oseltamivir resistance. This included hospitalized patients who were treated with at least one course of oseltamivir without improvement Hindiyeh et al. (2010). Most of the patients were immunologically suppressed as a result of various underlying diseases. All 31 patient samples were positive for pandemic (H1N1) 2009 virus; this was based on qRT-PCR assay as previously described Hindiyeh et al. (2010). Out of the 18 patients, 6 (33%) were found to be positive for the H275Y mutation by qRT-PCR. The oseltamivir-resistant virus was however first detected in a sample obtained after 9 days of oseltamivir therapy Hindiyeh et al. (2010). The first sample of patient 1 was tested as sensitive for oseltamivir on 30th July, 2009. It then and sensitive and resistant viruses were detected in the second sample of patient 1 when tested on 9th August, 2009. Sequence analysis of the neuraminidase genes was then amplified from the two samples and this confirmed the results of the qRT-PCR assays Hindiyeh et al. (2010). Similar results were obtained for patient 4. It is interesting to know that only oseltamivir-resistant viruses were detected in samples from patients 2, 3, and 5. The results were also confirmed by sequence analysis Hindiyeh et al. (2010). Hindiyeh et al. (2010) concluded that the report of two oseltamivir-resistant pandemic (H1N1) 2009 virus outbreaks, in the United Kingdom and the United States, was a strong reminder that countries should follow the WHO recommendations and develop a vigilant system for detecting oseltamivir-resistant viruses, stating that the qRT-PCR assay described in their research can be used to screen large numbers of clinical samples for drug resistance with a very short turnaround time of 3 to 4 hours and at a very low cost. In addition, the assay detects low quantities of either sensitive or resistant viruses in a mixed infection, however, sequencing fails to detect the minor virus type.

Meijer et al. (2007) stated that due to the influenza pandemic threat, many countries were stockpiling antivirals in the hope of limiting the impact of a future pandemic virus. Since resistance to antiviral drugs was probably going to significantly alter the effectiveness of antivirals, surveillance programmes to monitor the emergence of resistance were of considerable importance. An inventory was conducted by the European Surveillance Network for Vigilance against Viral Resistance (VIRGIL) in collaboration with the European Influenza Surveillance Scheme (EISS) to evaluate antiviral susceptibility testing by the National Influenza Reference Laboratories (NIRL) in relation to the national antiviral stockpile in 30 European countries that are members of EISS during the 2006/2007 influenza season Meijer et al. (2007).

According to Meijer et al. (2007), the influenza viruses quickly became resistant and were readily transmitted during therapy with adamantanes. Alao, in many parts of the world, circulating A(H3N2) viruses have become naturally resistant. In countries such as Cambodia, Vietnam and other parts of south-east Asia, it was reported that a high proportion of A(H5N1) viruses which were isolated from poultry were adamantane resistant. Although resistance has been detailed in clade 1 A(H5N1) viruses, it was not uniform through all lineages of A(H5N1) Meijer et al. (2007). For the NAI drugs, the emergence of resistance to seasonal influenza viruses [A(H3N2), A(H1N1) and B] was documented, with the highest frequency (18%) being reported in children following the oseltamivir treatment in Japan. Oseltamivir resistance in severe human A(H5N1) virus infections were also reported following therapy. However, the wide circulation of NAI resistant human strains was not been reported, and these could be detected at very low frequency in seasonal influenza surveillance programmes in countries which had high drug consumption Meijer et al. (2007).

The method employed by Meijer et al. (2007) for this research was the use of questionnaires at a VIRGIL-EISS laboratory workshop on influenza antiviral susceptibility testing techniques at the Health Protection Agency (HPA) in London between 3rd and 6th October, 2006 Meijer et al. (2007). The participants were representatives of 17 National Influenza Reference Laboratories (NIRLs) and were asked the following questions:

1. Does your country have a stockpile of influenza antiviral drugs?
2. If yes, which antiviral drugs?
3. Do you do any NAI susceptibility testing?
4. If yes, which tests: genotypic (virus nucleic acid analysis) or phenotypic (virus susceptibility analysis, that is, determination of 50\% inhibitory concentration values)?
5. Do you do any adamantane susceptibility testing?

6. If yes, which tests (genotypic/phenotypic)?

7. Do you plan to either introduce or extend NAI or adamantane susceptibility testing in the season 2006/2007?

The questionnaire was also sent out widely after the course by the EISS coordination center to all 40 EISS NIRLs for verification and for completion by the remaining 23 NIRLs. The answers to the questionnaires were processed at the EISS coordination center and the reports were created by laboratory and by country. For a country which had more than one NIRL the most positive answer was taken for the report by country. However, if the response to a question was: “available”, “if necessary"” or “possibly”, those results were interpreted in the analysis as no actual testing or no actual plans to introduce or extend testing Meijer et al. (2007).

Meijer et al. (2007) concluded that even though stockpiles of influenza antivirals were available in almost all EISS countries in Europe, the surveillance systems to track antiviral resistance which is necessary to support the use of the stockpiled drugs were not widely available. Through collaborative efforts of VIRGIL and EISS, the countries were being facilitated to develop antiviral susceptibility surveillance systems. This was to further strengthen the level of pandemic preparedness in Europe as an enhanced antiviral susceptibility monitoring capacity and capability will improve the ability of a country to deliver rapid information on the appropriate use of the stockpiled antivirals in case of an introduction of a new, possible pandemic, influenza A virus subtype Meijer et al. (2007).

Summary

In this chapter, the origin of influenza A, how it is transmitted, how severe it can become and how risky it can be to a certain group of people. We also discuss some methods that were used by others for detecting influenza A virus. Another review described the history of influenza A and how the different influenza A viruses are not similar. Hence, forming the basis for using H3N2 in comparison with H1N1 to ascertain the differences and also find potential targets. All these findings point in one direction; that is, the need to continually research and find newer and modern methods for disease detection and possible drug indications for such diseases. We also deduce from this chapter how useful and effective the system level analysis can be in finding the potential targets of influenza A, Davis et al. (2014), Garten et al. (2009), Fouchier et al. (2000) which are all modern and old research have all emphasized on the need to continue to find new indications of influenza A to find a drug of which the virus will not eventually become resistant to.

CHAPTER THREE

RESEARCH METHODS

Introduction

The main objective of this study is to identify potential targets of influenza A virus using a systems level based model. In this chapter, the materials used and methods employed in this study are discussed. As stated earlier, a systems level of analysis was performed in order to identify the potential targets of influenza A, based on many different mathematical procedures, ranging from information theory to probability theory, also, how clusters were identified, how candidate proteins were identified are all discussed in this chapter

Putative (influenza) drug targets and disease associated genes

The drug bank (http://www.drugbank.ca/) is a collective effort to bridge the `depth versus breadth' gap between clinically oriented drug resources and chemically oriented drug databases. The drug bank was first released in 2006, it was designed to serve as a comprehensive, fully searchable *in silico* drug resource that serves as a link for sequence, structure and mechanistic data about drug molecules with sequence, structure and mechanistic data about their drug targets, including biotech drugs Wishart and Knox (2008). Drug bank serves as clinically oriented drug encyclopedia, and provides detailed, and up-to-date, quantitative, analytic or molecular-scale information about drugs, drug targets and the biological or physiological consequences of drug actions. It also serves as a chemically oriented drug database, where provision for many built-in tools for viewing, sorting, searching and extracting text, image, sequence or structure data is made Wishart et al. (2008). A discovery platform known as DisGeNET (http://www.disgenet.org/) has been designed to address different questions that concern the genetic underpinning of human diseases, it is also one of the largest repositories of its kind currently available, it is made up of over 380000 associations between over 16000 genes and 13000 diseases. DisGeNET integrates expert-curated databases with text-mined data, and covers information on Mendelian and complex diseases. It also includes data from animal disease models. It is score based on evidence which supports prioritization of gene-disease associations. The database is an open access resource and is available through a web interface. It offers one of the most complete collections of human gene-disease associations; it also provides a precious set of tools which can be used to investigate the molecular mechanisms underlying diseases of genetic origin. It was designed to fulfill the needs of different user profiles, including bioinformaticians, biologists and health-care practitioners Pinero et al. (2015).

The prolonged and uncontrolled usage of drugs in treating viral diseases can cause multiple drug resistance. Influenza A is a viral infection and hence has the ability to mutate, this suggests that the drug used for treating influenza A must be changed for effective treatment since the virus becomes resistant. Meijer et al. (2009) also talk about an oseltamivir resistant H1N1 virus in their research work. This therefore calls for an identification of putative drug targets and the disease associated genes of influenza A. This is where the drug bank and DisGeNet comes into play. And to serve as a source for all data which is required for drug target identification and drug repositioning. Drug targets are proteins, compounds or drugs can be used to identify these target and disease associated genes are simply genes that have a connection with a particular disease. The protein SPLUNC1/BPIFA1 has been found to be associated with a number of other infections, respiratory infections especially. We will therefore try to find out if it is associated with influenza A (H1N1 and H3N2) in this thesis. The same protein was identified in the research conducted by Leeming et al. (2015) on gammaherpesvirus in the respiratory tract, and Liu et al. (2013) in their research on the virus contributes to pulmonary host defense against *Klebsiella pneumoniae* respiratory infection.

Human Functional Interaction Sources

There are different human functional interaction data sources. The human network used in this work was produced from the different data sources discussed in the following subsections. The unified score was used to score the network.

Search Tool for Retrieval of Interacting Genes (STRING)

As the name suggests, STRING represents an ongoing effort to put different types of protein-protein association evidence under one common framework. Such an approach gives several unique advantages such as (i) various types of evidence are mapped onto a single stable set of proteins thereby enabling comparative analysis; (ii) known and predicted interactions often partially complement each other, leading to increased coverage; (iii) integrated scoring scheme can provide higher confidence when independent evidence types agree; and (iv) mapping and transferring interactions onto a large number of organisms and enables evolutionary studies. STRING is fully pre-computed; therefore, all information can be quickly accessed, both at the high-level network view and at the level of individual interaction record. According to Von Mering et al. (2003), the database is an exploratory resource, it contains a much larger number of associations than primary interaction databases although the confidence scores vary. The protein-protein associations in STRING are imported from other databases such as conserved genomic neighbourhood, database imports (knowledge), phylogenetic co-occurrence, high-throughput experiments, and literature and co-expression analysis. STRING also contains a large body of predicted associations that are produced anew Von Mering et al. (2003).

STRING has a scoring scheme or a combined score between any pair of proteins. This score is often higher than the individual sub-scores, and this expresses increased confidence when an association is supported by several types of evidence. The database currently covers 9,643,763 proteins from 2,031 organisms.

Biological General Repository for Interacting Datasets (BioGRID)

BioGRID is a database that stores and distributes a comprehensive collection of physical and genetic interactions. It is a curated biological database of protein-protein interactions, genetic interactions, chemical interactions, and post transitional modifications. Data from BioGRID for the purpose of this research is not scored because the confidence level of data from BioGRID is high due to the fact that it is already curated Stark et al. (2006).

Intact Database

IntAct is an open source database tool kit for storing, presenting and analysing protein interactions Hermjakob et al. (2004). Protein interactions as we know provide an extremely useful resource for the explanation of cellular function. The IntAct data model is made up three components namely: Experiment, Interaction, and Interactor. The experiment groups a number of interactions, often from one publication and classifies the experimental conditions in which these interactions have been generated Hermjakob et al. (2004). According to Hermjakob et al. (2004), an experiment may have a single interaction, it may also have hundreds of interactions in the case of large scale experiments. They also explained that an interactor is a biological unit which participates in an interaction. It is usually a protein but has the potential to also be a DNA sequence or even a small molecule. Intact data is also curated manually; hence we set the confidence level of the data to be 0.8 which is high.

Database of Interacting Proteins

Database of Interacting Proteins (DIP) is a database that documents experimentally determined protein-protein interactions. The aim of DIP is to integrate the assorted body to experimental knowledge about interacting proteins into a single and easy to access database. They explain that DIP combines information from multiple observations and experimental techniques as well as providing information about networks of interacting proteins. Entries in DIP are done manually by the curator; automated tests are also performed to show that the proteins and citations exist. Interactions are also double checked by a second curator and flagged accordingly in the database (Xenarios et al. 2000). Hence, the reliability scores of data from DIP for this research is set at 0.8 which is high.

Sequence Data

Sequence data comprises of protein family and domains and sequence similarity. Protein sequence and domain signature datasets are collected from UniProt and InterPro databases respectively. The protein sequences provided by UniProt come from translation of coding sequences (CDS) which are submitted to different nucleotide sequence resources of the International Nucleotide Sequence Database Collaboration (INSDC), the coding sequences are generated by gene prediction programs or experimentally proven. The translated CDS sequences are then transferred automatically to the TrEMBL section of UniProtKB. InterProprovides functional analysis of proteins by classifying them into families and predicting domains and important sites. Protein signatures are combined from a number of member databases into a single searchable resource. Table 2 gives a summary of the data used and their sources.

Table 2 : Data Sources

Database	Description	Data type	Source
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	Pretreated protein interaction	http://string-db.org
BioGRID	Biological General Repository for Interaction Database	Physical and genetic interactions	http://www.thebiogrid.org
DIP	Database of Interacting Proteins	Protein interactions	http://dip.doembi.ucla.edu
IntAct	Molecular Interactions	Experimentally determined protein interactions	http://www.ebi.ac.uk/intact
UniProt	Universal Protein Resource	Protein sequence data	http://www.uniprot.org
InterPro	Integrated documentation resources for protein families, domains and functional sites	Protein signature or shared domain	http://www.ebi.ac.uk/interpro
Drug Bank	A unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target	Drugs and drug targets	http://www.drugbank.ca/drugs

Scoring Protein-Protein Functional Interactions

This section presents different scoring schemes for the different data collected specifically for this research work. Here, we describe how sequence data, BioGrid data, IntAct data, DIP data, and STRING data are scored for use in this research work. Apart from sequence data that is scored based on the scoring scheme explained in this chapter, every other dataset used is already scored based on the scoring schemes provided by the databases which the data is collected from.

Scoring Interactions from Sequence Data

Sequence data comprises of protein family and domains and sequence similarity. Protein sequence and domain signature datasets were collected from UniProt and InterPro databases respectively. We used information theoretic based approaches to derive scoring function models estimating confidence levels of functional interactions derived from these data. It is worth mentioning that sequence similarity data was used for scoring functional interactions from protein sequence data was obtained by using BLAST tool (Agustino, 2012).

Scoring Scheme for Protein Family and Domains

Two proteins a_i and a_j that share signature or entries $x_i, \dots x_m$ are considered. The similarity score, μ_{ij} of proteins a_i and a_j is defined as the minimum number of occurrences of the signatures in a_i and a_j . This is represented mathematically as follows

$$\mu \equiv \mu_{ij} = \sum_{k=1}^{m}(\mu_{ki}, \mu_{kj}), \tag{3.1}$$

where μ_{ki} is the number of occurrences of S_k in a_i, μ_{kj} is the number of occurrences of S_k in a_j. The confidence level C of the similarity score μ is given by

$$C \equiv C(\mu, \sigma, \alpha) = \emptyset\left(\frac{\mu^{\alpha}}{\sigma}\right) \tag{3.2}$$

where μ is the similarity score, σ is the dispersion measure, and α is the calibration control parameter and $\emptyset$ represents the cumulative probability of the standard Gaussian distribution defined by

$$\emptyset(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} exp\left(-\frac{x^2}{2}\right) \tag{3.3}$$

The Gaussian distribution was used because the data should have a normal distribution. A normal distribution makes the computation of the standard deviation also possible.

A training dataset D is a dataset made up of all pairs (S_k, Y_k) where, Y_k is the number of times the signature S_k is observed. A rectified dataset is however recommended to remove all outliers, in other words, observations that lie at abnormal distances from the data. The rectified dataset, D_r is a subset of the dataset D, and consists of a data point that falls inside $1.5 \times IQR$ represented as

$$D_r = \{(S_k, Y_k) \epsilon\, D: Q_1 - 1.5(IQR) \leq Y_k \leq Q_3 + 1.5(IQR)\}, \tag{3.4}$$

where Q_1 and Q_3 are the lower and upper quartile respectively, IQR, which is the interquartile range is computed as $IQR = Q_3 - Q_1$ and σ is the standard deviation of the rectified dataset estimated from maximum likelihood given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{1}^{N}(X_k - \bar{X})}^{2} \quad (3.5)$$

where N is the number of signatures in the rectified dataset and

$$\bar{X} = \sum_{k=1}^{N}\frac{X_k}{N} \quad (3.6)$$

is the mean of the set. The uncertainty measure related to the outcome of μ resulting from the data was obtained from the binary entropy function is given by

$$H_2(\delta) = -\delta \log_2(\delta) - (1-\delta)\log_2(1-\delta) \quad (3.7)$$

The uncertainty measure of the function $H_2(\delta)$ is defined in the interval[0,1],$H_2(0) = H_2(1)$since

$$\lim_{s\to 0^+} s\log_2(s) = 0 \quad (3.8)$$

and

$$\lim_{s\to 1^-}(1-s)\log_2(1-s) = 0 \quad (3.9)$$

The capacity of inferring the functional relationship score between two proteins belonging to the same family or sharing common signatures is given as

$$\Gamma(\delta) = 1 - H_2(\delta)$$

The reliability or confidence score of the functional relationship between two proteins is given by

$$R = \frac{\Gamma(\delta)}{\max_s \Gamma(s)}$$

When μ is significantly large, δ converges to 1, hence, the uncertainty measure of $H_2(\delta)$ converges to 0. This leads to the maximum capacity of inferring the functional relationship of 1 and hence the reliability of the functional relationship between two proteins is given by

$$R = \frac{\Gamma(\delta)}{bit}$$

Scoring scheme for protein sequence similarity

Here, the bit score alignment of homologous sequences s_1 and $s_2$represented by $S(s_1, s_2)$ is set with its standard units and $I(s_1, s_2)$ is the mutual information between the two sequences. Hence, the bit score is represented as

$$S(s_1, s_2) = \delta \times I(s_1, s_2),$$

where γ is a constant defining the unity.

In a general sense, $S(s_1, s_2) \neq Ss(s_2, s_1)$ and will be equal only if they have the same scale for the search size. On the other hand, the mutual information$I(s_1, s_2)$between two sequences s_1 and s_2 satisfies

$$I(s_1, s_2) = I(s_2, s_1) \qquad (3.10)$$

and

$$I(s_1, s_2) \geq 0 \qquad (3.11)$$

The reliability score or confidence score$R(s_1, s_2)$ of a functional relationship between two protein sequences s_1 and s_2 is defined as normalised mutual information and was calculated as

$$R(s_{1,}s_2) = \frac{I(s_1,s_2)}{max\{H(s_1),H(s_2)\}} \qquad (3.12)$$

This measured how the protein sequence s_1 was able to predict the protein sequences_2, where $H(s_1)$ is the relative entropy obtained by aligning a protein sequence s by itself.Once mutual information increases with relative entropy, this yields a bias, and this bias is corrected by a division of the mutual information by the maximum entropy of the sequence pair. The mutual information $I(s_1, s_2)$ can therefore be computed as

$$I(s_1, s_2) = \frac{S(s_1,s_2)+S(s_2,s_1)}{\lambda+\lambda\prime} \tag{3.13}$$

where λ and λ' are constants that define unity for $S(s_1, s_2)$ and $S(s_2, s_1)$ respectively. Also, for a protein $s, H(s) = I(s, s)$ is given by

$$H(s) = \frac{2\times S(s,s)}{\lambda+\lambda\prime} \tag{3.14}$$

Substituting Equations (\ref{e3}) and (\ref{e2}) into (\ref{e1}), we obtain the reliability and confidence score between two protein sequences, which is given by

$$R(s_1, s_2) = \frac{S(s_1,s_2)+S(s_2,s_1)}{2\times max\{S(s_1,s_1),S(s_2,s_2)\}} \tag{3.15}$$

This scoring scheme is independent of constant λ and λ', and only depends on the two proteins for which the confidence score is being computed. If the similarity score of two proteins of which the mutual information of their evolutionary history is embedded is 0, it means the two sequences are not similar and therefore, their confidence score is also 0.

Scoring Interactions from STRING Database and Others

STRING is a database of known and predicted protein-protein interactions.

The interactions include direct physical and functional associations. These come from computational prediction, knowledge transfers between organisms, and from interactions aggregated from other databases Von Mering et al. (2003). STRING interactions are derived from five main sources namely: genomic context predictions, high-throughput lab experiments, conserved co-expressions, automated text mining, and previous knowledge in databases. The scoring scheme for STRING data is given by

$$S = 1 - \prod_i (1 - s_i) \tag{3.16}$$

where i is the score from each data source for STRING.

All entries in DIP are done manually by the curator, automated tests are also performed to show that the proteins and citations exist, since interactions are double checked by a second curator and flagged accordingly in the database, we set the reliability scores of data from DIP Xenarios et al. (2000) for this research is set at 0.8. Similarly, for Intact Hermjakob et al. (2004) and BioGRID datasets Stark et al. (2006), the confidence level is set at 0.8 due to the fact that, data from the Intact and BioGRID databases are manually curated and double checked.

Unified Interaction Scoring Function and Effectiveness

The unified scoring system

After cleaning and extracting the required STRING data which is already scored, and the scored sequence data, a unified score for a unified network is created by combining already scored data from BioGRID, DIP and IntAct in addition to that of the STRING and Homology data, the Unified score is computed using the computation

$$S_{ij} = 1 - \prod_{d=1}^{n}\left(1 - s_{ij}^{d}\right), \qquad (3.17)$$

where n is the number of data sources considered, d the data type, and s_{ij}^{n} is the confidence score of functional interaction between proteins i and j derived from the source of the data type d.

Effectiveness of the scoring function

The confidence score of a functional interaction between proteins p and q measures the confidence level in this specific functional interaction and represents the probability of the occurrence of this interaction. First assume that n different sources were used to predict this interaction and let$\overline{A_{pq}}$ be an event indicating that the functional interaction between proteins p and q could not be inferred from any of these n sources under consideration, that is,

$$\overline{A_{pq}} = \bigcap_{d=1}^{n} \overline{A_{pq}^{d}} \qquad (3.18)$$

with $\overline{A_{pq}^{d}}$ being the event indicating that the functional interaction could not be retrieved using the source d.

Under the assumption that sources are independent, the probability $\mathrm{P}\left(\overline{A_{pq}}\right)$of the event $\overline{A_{pq}}$ is given by

$$\begin{aligned}\mathrm{P}\left(\overline{A_{pq}}\right) &= \mathrm{P}\left(\bigcap_{d=1}^{n} \overline{A_{pq}^{d}}\right)\\ &= \prod_{d=1}^{n} \mathrm{P}\left(\overline{A_{pq}^{d}}\right) \qquad (3.19)\\ &= \prod_{d=1}^{n}\left(1 - \mathrm{P}\left(A_{pq}^{d}\right)\right)\end{aligned}$$

where A_{pq}^{d} is the event indicating that the functional interaction is retrieved using the source d and thus $\mathrm{P}\left(A_{pq}^{d}\right) = s_{pq}^{d}$ with s_{pq}^{d} the confidence score of a functional association between p and q predicted using the source d. Thus, the combined confidence score S_{pq} for interacting proteins p and q, which is the probability of the event A_{pq}, which indicates that the functional interaction between proteins p and q can be inferred from at least one of the sources, contrary to $\overline{A_{pq}}$, is given by

$$\begin{aligned} S_{pq} &= \mathrm{P}\left(A_{pq}\right) \\ &= 1 - \mathrm{P}\left(\overline{A_{pq}}\right) \qquad (3.20) \\ &= 1 - \prod_{d=1}^{n}\left(1 - \mathrm{P}\left(A_{pq}^{d}\right)\right) \end{aligned}$$

It follows that

$$S_{pq} = 1 - \prod_{d=1}^{n}\left(1 - s_{pq}^{d}\right) \qquad (3.21)$$

Finally, an illustration on how other scoring functions, such as minimum (min), maximum (max) and average (mean) of different confidence scores may produce biased combined or unified score, and why the scoring function in the equation $S_{pq} = 1 - \prod_{d=1}^{n}\left(1 - s_{pq}^{d}\right)$ is more realistic. Assume that out of $n = 11$ different data sources, the functional interaction between proteins p and q was predicted from 2 sources out of 11 with confidence scores of 0.200 and 0.130. So, for any other source, the confidence score is assumed to be 0, and it follows that,

Using themin function, we get,

$S_{pq} = \min\{0, 0, 0, 0, 0, 0, 0, 0, 0, 0.200, 0.130\}$, which implies that $S_{pq} = 0.00$ indicating that the confidence score is 0 and this interaction will be ignored in different analyses whereas it was predicted by two different sources.

Using max and min, the combined confidence score, S_{pq}, is equal to 0.200 and 0.030. The max function does not reflect the fact that the functional interaction was predicted from two different sources and the mean function reduces our confidence level. As this interaction was predicted by two different sources, by instinct, it is expected that the confidence level will increase, instead it is decreasing. This suggests that these scoring functions are not in agreement with what can be expected and show biases by underestimating combined interaction scores in the final network. On the other hand, using the scoring function in the equation $S_{pq} = 1 - \prod_{d=1}^{n}\left(1 - s_{pq}^{d}\right)$ as used in this research work, we have $S_{pq} = 0.304$, this shows a more realistic combined confidence score compared to other scoring functions, and this agrees with what is expected.

Gene Ontology Annotation and Pathway Dataset

Gene Ontology (GO) is a major collaborative bioinformatics initiative.

Its aim is to standardize the representation of gene and gene product attributes from species to species. The project provides a controlled vocabulary of terms for describing gene product characteristics, supports gene product annotation data from GO Consortium (GOC) members, and develops tools to access and process these data. For the past decade, the GOC has expanded from its three model organism databases (mouse, yeast and fly). It now includes the world's major repositories for plant, animal and microbial genomes. GOC makes its ontologies, annotations, and tools freely available to advance biological research Consortium (2010). The Gene Ontology presents three aspects of biology namely molecular functions (MF), biological processes (BP), and cellular components (CC). The GO project aims to specifically:

1. Maintain and develop its controlled vocabulary of gene and gene product attributes;
2. Annotate genes and gene products, and assimilate and disseminate annotation data; and
3. Provide tools for easy access to all aspects of the data which is provided by the project, and to allow the functional interpretation of experimental data using the GO. An example of such usage is enrichment analysis.

GO annotations are statements that describe the functions of specific genes by making use of concepts in GO. The most basic and common annotations link one gene to one function. Annotation is also a process of assigning GO terms to gene products (www.geneontology.org). The dataset for GO Annotation (GOA) can be found at the GOA database. The GOA database is the source of high-quality electronic and manual annotations to the UniProt knowledge base. GOA can also be used as a source of solution for biological problems Camon et al. (2004).

Pathway is a kind of map that describes the biological mechanism of an organism. Diseases are identified with genetic modules such as pathways. Pathway level analysis gives important insights when it comes to making biological inferences and hypothesis from genetic data Guo (2010). There have been few approaches for incorporating biological pathways knowledge in the interpretation of high-throughput datasets until recently. ``Biological pathways represents the biological reactions and interaction network in a cell" Guo (2010). The evidence that the connection between pathways and diseases is constructed at many interconnected levels and is controlled by an interaction between cell signaling, gene expressions and so on is becoming abundant. Identification and alignment of pathways are very useful for the inference of the biological mechanism of diseases from high-throughput genetic data. There are different pathway databases, one of which is the Kyoto Encyclopedia of Genes and Genomes (KEGG) which includes the topology of pathway Guo (2010). Most of the different pathway databases often make use of a different data structure and terminology, hence making the discovery of conserved pathways between different species complex. Gene Ontology (GO) however provides an ordered structure of concepts on molecular function, biological processes, and cellular component. When genes in a pathway are mapped to their ontology terms using the semantic structure of GO in order to define the similarity of the genes gives pathway alignment a basis Guo (2010). In this research work, KEGG pathways are used.

Network Topological Structure Analysis

A network can be described as a set of nodes connected by interactions. A network can represent traffic between airports, or the social interaction between workers in an office. Analysing complex networks in science is known as network theory Barrenäs (2012). The subject of network theory has already been discussed in Chapter One. Network theory is a tool for systems biologists use, disease-associated genes can be organised through different biological sources such as protein-protein interactions, co-expressions and so on to form networks. These networks are then analysed based on their topological structure to detect disease pathways, clinical markers, and drug targets Barrenäs (2012).

Network centrality measures

Centrality is a way of placing priorities on nodes within a network. When a node is well connected, with the rest of the network, it is considered as a central node. This implies that removing this node from the network affects the unity of the network. Network centrality can be used to detect disease-associated genes. This is because such genes usually have a higher centrality. Disease-associated genes with high centrality are important for the organism Barrenäs (2012).

Degree is a local centrality measure, there are also two common global centrality measures also known as closeness and betweenness. Degree centrality corresponds with the number of interactions or immediate neighbours a node has Barrenäs (2012). The degree of a protein shows the influence it has on a biological process occurring in an organism, this implies that a protein with more functional connections tends to contribute to several processes and is likely to be a key protein in the functioning of the system (Mazandu & Mulder, (2011). The *Degree* of a protein p is the number of links connected to it, that is, the number of its interacting neighbours and is represented as

$$\deg(p) = \sum\nolimits_{q \epsilon N} \delta(p,q) \tag{3.22}$$

N is number of proteins in network and

$$\delta(p,q) = \begin{cases} 1, & if\ q\ is\ linked\ to\ p \\ 0, & otherwise \end{cases}$$

The *Betweenness* centrality of a protein p in a functional network is a metric that expresses the influence of p relative to the other proteins in the network. This is built on the proportion of the shortest paths between other proteins which pass through the protein target. It shows the importance of a protein for transmitting information between other proteins in the network (Mazandu & Mulder, (2011)). The betweenness measure helps to identify the number of pairwise proteins connected indirectly by the protein target through their direct functional connections. The proteins with high betweenness are expected to ensure the connectivity between proteins in the functional network, these proteins are also able to bridge or disconnect connected components. In other words, nodes that have many short paths passing through them have a high betweenness. The betweenness is calculated by identifying the shortest paths between all the nodes within the network first Barrenäs (2012), and represented mathematically as follows

$$B(p) = \sum\nolimits_{(s,t) \epsilon N_p} \frac{\sigma_{st}(p)}{\sigma_{st}}, \tag{3.23}$$

where $\sigma_{st}(p)$ is the number of shortest paths from protein s to protein t passing through p, and σ_{st} is the number of shortest paths from s to t in the functional network.

The *Closeness* shows the ability of a protein to access information through other proteins and to transmit information in the network. The closeness measure $C_s(p)$ of a protein p is the inverse of its status. The status, $S(p)$ of a protein p in a connected network is the average distance to all other proteins, that is, the ratio of the sum of $\pi(p,q)$ for all proteins q in the network to the total number of such paths is $(n-1)$. The status is given by

$$S(p) = \frac{1}{(n-1)} \sum_{q \epsilon N} \pi(p,q). \tag{3.24}$$

In summary, closeness measures the average distance to all other nodes (proteins) and represented mathematically as

$$C_s(p) = \frac{|L_c|-1}{(n_c-1) \times S_{r(p)}}, \tag{3.25}$$

where $S_r(p)$ is the status of p relative to its connected component. The above closeness can be normalized by $\frac{(n_c-1)}{|L_c|-1}$ to account for an instance where the functional network is not completely connected. The number of nodes in the connected part of the network is given by n_c and the number of functional links in the connected component is given by L_c. The normalization makes the scale uniform for comparison.

Protein Degree and Path Length Distribution

The degree is a key property of a node. It represents the number of links a node has with other nodes. The degree can represent the number of contacts an individual has on his phone in a call network, it can also represent the number of proteins that are connected to a particular protein in a protein interaction network Barabási (2016). The degree distribution of a protein is denoted by p_k, where p_k is a fraction of nodes in the network that have degree k For a random graph, each edge is present or absent with equal probability, hence making the degree distribution Binomial or Poisson according to a study by Erdõs and Rényi (Newman, 2003b). Real world networks however, seem to be different in their degree distribution, unlike the random graphs which were used byErdõs and Rényi. The degree distribution of nodes in most networks are highly rightly skewed, implying that their distributions have a long right tail which are far above their mean (Newman, 2003b).

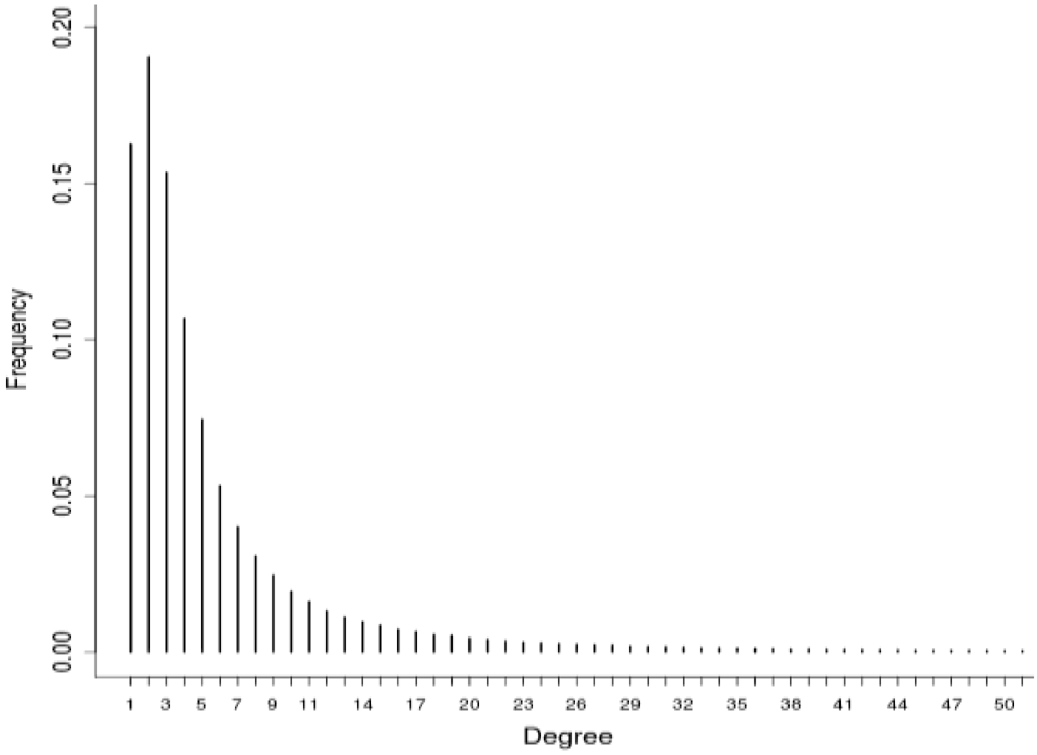

Figure 1: Degree Distribution Illustrated Graphically

Figure 1 illustrates the skewed to the right tail of a degree distribution. Measuring the tail can be dicey, theoretically, a histogram of the degrees can be constructed to measure this, however, it cannot be accomplished practically because there are rarely enough measurements to get good statistics in the tail, direct histograms are often noisy, where noisy refers to fluctuations of data that makes it more difficult to perceive the real picture or expected results (Newman, 2003b). There are networks which also have degree distributions that are approximated by power law, these networks are sometimes called scale-free networks. The degree distribution is represented mathematically as

$$P_k = \frac{N_k}{N}, \tag{3.26}$$

where N_k is the number of degree k nodes, for a network with N nodes. The number of degree k nodes can be found from the degree distribution as $N_k = NP_k$. Studying several dynamic processes over real networks has made known the existence of short cuts, that is, to bridge the links that connect different areas of a network, thus, speeding up the communication among distant nodes. In a regular hypercube lattice, in dimension D, the mean number of vertices that need to be passed by in or to reach the arbitrary chosen node grows with the size of the lattice as $N^{\frac{1}{d}}$ (Boccaletti et al., 2006). Centrality, most of the real networks, irrespective of their large size, have a relatively short path between any two nodes. This feature is known as the *small world property* and presented mathematically by an average path length L (Boccaletti et al., 2006). The average path length is computed as follows

$$L = \frac{1}{N(N-1)} \sum_{i,j\, \epsilon N, i \neq j} d_{ij} \tag{3.27}$$

This property is seen in the computing of closeness and betweenness in networks and is applied in this research work.

Identification of Network Key Proteins

Key proteins are very useful in systems level analysis. Just as the name, key proteins play a key role within the human. They do not change and hence identifying them and using them for drug repositioning is very important and useful. Identifying key proteins of the pathogen are not useful because the pathogen can mutate, a typical example is viruses, if such key proteins are identified today, they can change the next time the information. Hence, in this research work, we make use of key proteins from the human network.

The key proteins of the network are identified using the network centrality measures. From the Unified human network, the degree, closeness and betweenness are computed using the methods already discussed in this chapter under clustering. For a protein to be considered as a key protein, its closeness must be greater than or equal to $\frac{1}{\pi_p}$, the betweenness must be greater than or equal to $n \times \pi_p$; where n is the network size and π_p is the shortest path average, and the degree must be above the average degree. All proteins meet the degree and betweenness or degree and closeness criteria are considered as key proteins satisfy each measure, not just one or two.

Clustering Proteins in the Network

Clustering, which can also be called transitivity, is a typical property of acquaintance networks, that is, two individuals have a common friend and hence, likely to know each other. In terms of a generic graph G, transitivity implies that a high number of triangles is present. This can be quantified by defining the transitivity T of the graph as the relative number of transitive triples, that is to say, the fraction of connected triples of nodes (triads) which also form triangles Boccaletti et al. (2006). Another possibility is to use the graph clustering coefficient C, a measure introduced by Watts and Strogatz is defined as follows. A quantity c_i which is the local clustering coefficient of node i is first introduced, expressing how likely $a_{jm} = 1$ for two neighbours j and m of node i. This value is obtained by counting the actual number of edges which is denoted by e_i in G_i, the sub-graph of neighbours of i Boccaletti et al. (2006). Clustering or partitioning of a network is also considered when dealing with a huge network. To find the exact optimal partitions in network is known to be hard to computationally manage, this is because, of the explosion of the number of possible partitions as the number of nodes increases. There however have been different types of community detection of algorithms such as the divisive algorithm with which detects inter-community links and remove them from the network, agglomerative algorithms which merges similar nodes or communities recursively, there is also the spectral methods which are based on eigenvectors of the Laplacian matrix, there is also the optimization methods which are based on the maximization of a benefit function Blondel et al. (2008). The quality of resulting partitions from these methods from the methods listed above is usually measured by the modularity of the partition.

Blondel et al. (2008) introduced an algorithm that finds high modularity partitions of large networks in a short time, it also unfolds a complete hierarchical community structure for the network, by providing a clear description of this phenomenon and how useful their algorithm is in the partitioning of a network. This algorithm by Blondel et al. (2008) is implemented in this research work to cluster to the protein-protein networks which were created. The algorithm presented here finds high modularity partitions of large networks in short time and that unfolds a complete hierarchical community structure for the network. This therefore gives access to distinct resolutions of community detection. As opposed to most of the other community detection algorithms, the network size limits that was used with the Blondel et al. (2008) algorithm was due to limited storage capacity not limited computation time: it took only 152 minutes to identify 118 million nodes network. The algorithm was divided into two phases and repeated iteratively Blondel et al. (2008). A portion of the algorithm's efficiency is as a result of the fact that the gain in modularityΔQ,which is obtained by moving i in the community C of $j$was easily computed by

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,n}}{m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \qquad (3.28)$$

Estimating Protein Closeness at the Functional Level

We discuss different term information content models, term semantic similarity approaches and functional similarity measures between proteins in GO in this section. From differences (Mazandugo (2016), it is assumed that IC models are partitioned into two families namely, annotation based and topology based IC. Term semantic similarity are also split into 3 main categories, these are edge or path-based, IC or node-based, and hybrid. Functional similarities also have two man classes, these are: ontology and non-ontology based measures. It is also assumed that GO has three separate ontologies which have already been discussed in Chapter Three, these are: molecular function, biological process and cellular component. These also act as roots for the ontologies, hence the specificity cannot be found based on the root. Pathways will also be discussed in this session.

Computing IC Values

Since the inception of term information content (IC), the approaches can be divided into two families namely: annotation-based and topology-based IC families. While the topology-based family explores only the intrinsic topology of the GO directed acyclic graph (DAG), the annotation-based family requires an additional annotation data. Apart from the proposed Wang et al topology-based model, all other approaches compute the IC of terms in a similarly, that is, by using log function) despite their conceptual differences (mazandugo (2016). The IC value of the term is given by $\mathrm{IC}(x) = -\ln\big(p(x)\big)$. This section explores some of these approaches.

Annotation based model

In annotation-based approaches, $p(x)$ is the relative frequency of the term x in the protein dataset being considered. This is obtained from the frequency $f(x)$ and it represents the number of proteins, $\eta(x)$ annotated with the term x in the dataset. The "true-path rule" principle of the GO DAG structure is considered Mazandugo (2016). The frequency $f(x)$ is given by

$$f(x) = \begin{cases} \eta(x), & \text{if } x \text{ is real} \\ \eta(x) + \displaystyle\sum_{z \in C_h(x)} \eta(x), & \text{otherwise} \end{cases} \tag{3.29}$$

where $C_h(x)$ is the set of GO terms having x as parent, and a leaf is a term with no child.

Topology based models

GO-universal IC in terms of GO-universal approach, $p(x)$ is called the topological position characteristic of x, which is recursively obtained by using its parents that are gathered in the set $P_x = \{t: (t, x) \in L_{GO}\}$ where L_{GO}, expresses the set of links in the GO-DAG and$(t, x) \in \ L_{GO}$ is a representation of the link or association between a given parent t and its child x. The topological position characteristic, $p(x)$, is denoted by

$$P(x) = \begin{cases} 1 & \text{if } x \text{ is a root} \\ \prod_{t \in p(x)} \frac{P(t)}{|C_h(t)|}, & \text{otherwise} \end{cases} \tag{3.30}$$

There is other topology based methods that are equally useful for finding GO similarity scores. Some of these are: Zhang et al. (2006), who proposed a topology based approach where $f(x)$ is called the count of the term x and is dependent on only the children of the given GO term. This is numerically equal to the sum of counts of all its children. In this method, the similarity is computed using a recursive formula, starting from the leaves in the hierarchical structure. Other people such as, Seco et al. (2004), Zhou et al. (2008), Seddiqui & Aono (2010), and so on all proposed topology based IC models that can also be used for computing the similarity scores of proteins in the GO structure Mazandugo (2016).

GO term semantic similarity approaches

There are also term semantic similarity approaches namely, node, and edge approaches. For IC-based or node-based GO term semantic similarity, several approaches have been proposed for computing the scores within the GO

Directed Acyclic Graph (GO-DAG). Some of the approaches are Resnik, Lin, and many others which will be discussed in this section. Also, some new approaches such as Aggregate information content (AIC), relevance similarity by Schlicker et al. (2006), and information coefficient similarity by B. Li et al. (2010) were proposed to improve existing GO term comparison according to (Mazandugo, 2016).

The Resnik, Lin, Nunivers, FaITH and P&S approaches

For the Resnik approach, computes the similarity between two terms with the information content of the most informative common ancestor (MICA) of the two terms and this is computed as

$$S_r(a,b) = \mathrm{IC}(c) = \max\{\mathrm{IC}(x): x \in A_a \cap \ A_b\}, \qquad (3.32)$$

where c is the most informative common ancestor between the two terms, a and b.

The Lin approach for computing semantic similarity also makes use of the MICA between the terms that are under comparison, it is also normalised by the values of the IC terms. The Lin approach can produce scores between 0 and 1 which satisfies the property of having a semantic similarity score that lies between a term and itself is 1. It was however not the same with the Resnik approach. The semantic similarity between two terms using the Lin approach is given by

$$S_l(a,b) = \frac{2\times \mathrm{IC}(c)}{\mathrm{IC}(a)+\mathrm{IC}(b)} \qquad (3.33)$$

In order to have the scores between 0 and 1, there were two schemes were proposed. The first was using the possible upper bound of IC values which was called the Nunif and the other was using the highest possible score in the ontology under discussion which was called the Nmax. The two schemes are represented by

$$S_r(a,b) = \begin{cases} \frac{\mathrm{IC}(c)}{\log_2 N} \text{for Nunif} \\ \frac{\mathrm{IC}(c)}{\mathrm{IC}_{\max}} \text{for Nmax} \end{cases}$$

where N is the number of annotated proteins in the collection under study and IC_{max} is the highest IC score in the ontology considered.

Another approach known as the Nunivers was proposed to complete the requirement of a semantic similarity score which is that the score between a term and itself should be 1 by normalizing the score using the maximum IC values of terms which is given by

$$S_n(a,b) = \frac{\mathrm{IC}(c)}{\mathrm{IC}(a)+\mathrm{IC}(b)+\mathrm{IC}(c)} \qquad (3.34)$$

Again, the FaITH approach which is an adaptation of the edge-based semantic similarity which was introduced by Stojanovic et al is computed as

$$S_{FaITH} = \frac{\mathrm{IC}(c)}{\mathrm{IC}(a)+\mathrm{IC}(b)+\mathrm{IC}(c)} \qquad (3.35)$$

For the P&S approach, they suggested that the semantic similarity score should be computed as

$$S_{\mathrm{P\&}S}(a,b) = \begin{cases} 1 & \text{if } a = b \\ max\{3 \times \mathrm{IC}(c) - \mathrm{IC}(a) - \mathrm{IC}(b)\}, & \text{otherwise} \end{cases}$$

This is defined this way to prevent the violation of the non-negativity property of semantic similarity measures.

Edge-based GO Term Semantic Similarity

The edge-based approach according to (Mazandugo, 2016) is the oldest approach for measuring semantic similarity between terms in a hierarchical structure that was proposed. The semantic similarity in this case is a function of the number of edges on a path between two terms. Two terms are more semantically similar depending on the shortness of the shortest path length between them. Some approaches were also proposed and we will take time to briefly discuss them.

The approach by Rada et al. is as follows. For two terms under consideration, their approach was dependent on the shortest distance in terms of the number of edges. The semantic similarity between two terms a and b is therefore the multiplicative inverse of the length of the shortest path between the terms. To prevent division by zero, 1 is added. The computation is given by

$$S_{rada}(a,b) = \frac{1}{1+D_{sp}(a,b)} \tag{3.36}$$

Resnik also proposed an edge-based approach, where the shortest distance between a and b is converted to a semantic similarity by subtracting the shortest distance from the maximum path length in the ontology computed as

$$S_{re}(a,b) = 2 \times \partial_{\max} - D_{sp}(a,b) \tag{3.37}$$

where $D_{sp}(a,b)$ is the shortest distance. To obtain a normalized version of Resnik's edge-based approach, each term is divided by Equation 3.37 using the possible maximum value, which is $2 \times \delta_{\max}$ Therefore, the normalised Resnik edge-based semantic similarity measure is given by

$$S_{nre}(a,b) = 1 - \frac{D_{sp}(a,b)}{2\times\delta_{\max}} \qquad (3.38)$$

Leacock and Chodorow also proposed an approach to finding semantic similarity scores. Their approach is similar to the Resnik edge-based approach. They however introduced the use of a non-linear function 'log' to convert the shortest distance $D_{sp}(a,b)$ to semantic similarity score $S_{lc}(a,b)$. Hence, the similarity score between two terms is the negative log of the ratio between the length of the shortest path and two times the maximum depth of the hierarchy under consideration(Mazandugo, 2016).

There are several other ways of finding similarity scores, however, a few are highlighted in this work. The GO universal and node based approached are specifically used in this research work for finding GO-universal similarity is chosen because, it is the appropriate tool for scoring the term specificity in the GO DAG Mazandu & Mulder, (2013b).

Functional similarity measures

A protein can be annotated by a set of terms because it can perform more than one biological function also be involved in several processes. A functional similarity can be measured between sets of GO terms annotating proteins, an example is the use of the IC values of their GO terms which is directly referred to as group-wise measures, it is also known as direct term score based measure. One can also use term semantic similarity scores of pairwise GO terms between GO terms as an alternative, and this is known as pairwise or term semantic-based measure (Mazandugo, 2016).

Several measures for estimation of scores in terms of annotation-based IC approaches have been proposed to enable protein comparisons at the functional level. These functional similarity scores are obtained by using statistical measures of closeness, such as average (Avg), maximum (Max), best-match average (BMA) and averaging all the best matches (ABM) (Mazandugo, 2014). The average and maximum measures are computed as

$$Avg(p,q) = \frac{1}{n \times m} \sum_{s \in T_p^X,\ t \in T_q^X} S_{GO}(s,t) \qquad (3.39)$$

and

$$\text{Max}(p,q) = \max\{S_{GO}(s,t): s \in T_p^X \text{and } t \in T_q^X\} \qquad (3.40)$$

where T_r^X is a set of GO terms in X representing the molecular function (MF), biological process (BP) or cellular component (CC) ontology which annotates a given protein r, and $n = |T_p^X|$and $m = |T_q^X|$are the number of GO terms in these sets, and$S_{GO}(s,t)$ is the semantic similarity score. The ABM is the mean of best matches of GO terms for two annotated proteins, each against the other, given by the following formula (Mazandugo, 2014)

$$\text{ABM}(p,q) = \frac{1}{n+m}\left(\sum_{t \in T_p^X} S_{GO}\left(t, T_q^X\right) + \sum_{t \in T_q^X} S_{GO}(t, T_p^X)\right) \qquad (3.41)$$

with $S_{GO}\left(t, T_p^X\right) = \max\{S_{GO}(s,t): t \in T_r^X\}$. The Best Match Average (BMA) for two annotated proteins $p$and $q$is the mean of two values, the first is the average of best matches of GO terms annotated to protein p against those annotated to protein q, and the second is the average of best matches of GO terms annotated to protein q against those annotated to protein p, given by the following formula (Mazandu, 2014)

$$\text{BMA}(p,q) = \frac{1}{2}\left(\frac{1}{n}\sum_{t \in T_p^X} S_{GO}\left(t, T_q^X\right) + \frac{1}{m}\sum_{t \in T_q^X} S_{GO}(t, T_p^X)\right) \qquad (3.42)$$

The four functional similarity measures described above require GO term semantic similarity scores, and are also known as IC-based term semantic similarity based measures (Mazandu, 2014). For the topology-based family, each approach has been suggested with its functional similarity measure. The GO-universal metric however uses the BMA, the ABM was used in the Wang et al. approach. The Zhang et al. measure is a context dependent approach and authors initially suggested using the approach proposed by Lord et al., which is the Avg scheme for measuring functional similarity scores between proteins (Mazandu, 2014). There are also annotation based measures which involve measuring the semantic similarity of two GO terms based only on the most informative common ancestor terms (Mazandu, 2014). In this research work, we use the BMA measure, this is because this approach considers all terms of the proteins, hence, there is no loss of information, it compares only each term with its most similar therefore, it is not biased by the number of annotations per protein (Pesquita, 2008).

Extraction of Differentially Abundant Proteins

Differentially abundant proteins in samples from IAV positive patients versus IAV negative patients were analysed at the systems level to identify final set of targets using several criteria, which included statistical and biological relevance. We then filtered the set of differentially abundant proteins using their abundance levels and checking whether they have been previously identified. We then mapped the proteins to the generated human protein-protein functional network to elucidate their roles in the system based on network centrality measures. The extraction process is then started by ranking the differentially abundant proteins in samples using their abundance-level statistics. The ranked gene list is used to compute a Pearson Chi-Square Score χ_P^2 for every single gene subset, this shows the tendency of genes in a particular set to occur towards the extremes of the list. The PS of a subset containing proteins re-indexed from n to m is given by

$$\chi_p^2(n,m) = \sum_{i=n}^{m} \frac{(x_i-\mu_{nm})^2}{\mu_{nm}}, \qquad (3.43)$$

where x_i is the abundance score of protein at i, andμ_{nm} is their expected value. The χ_P^2 is known to approximate the χ^2-distribution with$n-m$degree of freedom (dof) which has a of variance $2 \times (n-m)$. The expected value can be computed as

$$\mu_{nm} = \frac{1}{m-n+1} \sum_{i=n}^{m} x_i \qquad (3.44)$$

Here, s top proteins were identified as differentially higher abundant proteins with s as the index fold-change χ^2. Assuming that the ranked list contains S proteins,$s(s \leq S)$is the smallest index which satisfies the following inequality

$$S = \frac{\chi_P^2(1,s) - r(\chi_P^2(s+1,S)}{\sqrt{1+r^2}} \qquad (3.45)$$

where

$$r = \sqrt{\frac{s-1}{S-s-1}} \qquad (3.46)$$

We set $r = 1$, if $s = 1$, r is the ratio of standard deviations of $\chi_P^2(1, s)$ and $\chi_P^2(s + 1, S)$. Therefore, the Estimated Score(ES) of the extracted subset is calculated as

$$ES(S) = \sum_{i=1}^{s} x_i \qquad (3.47)$$

The significance of estimated score of the subset of differentially higher or lower abundant proteins is assessed using sample randomization. Hence, we randomly selected 1000 independent subsets (of same size s) of differentially abundant proteins and compute ES of each subset and then perform the Shapiro-Wilk test under the null hypothesis that the generated sample is drawn from a normal distribution. Based on the rejection or acceptance of the null hypothesis, a Wilcoxon or T-test is performed to check whether the identified set of differentially higher or lower abundant proteins is more than expected by chance.

Filtering the Set of Differentially Abundant Proteins

Differentially abundant proteins in samples from IAV positive patients versus IAV negative patients were subjected to two filters in order to be considered as targets. For a protein to be a target, it should be:

1. Either on the set of differentially higher or lower abundant proteins or previously identified, that is, it should be found in the list of IAV disease associated genes extracted from the DisGeNet database, and
2. A key protein in the human protein-protein functional network, that is, having the network centrality scores (degree, closeness and betweenness) reaching the threshold.

A human protein-protein functional network was built using the protein-protein interaction datasets extracted from several sources already discussed in this work, including the STRING database, homology datasets (sequence similarity and shared InterPro domain signatures), the BioGRID database, IntAct, and DIP databases. These interacting datasets were used to form a single unified network with a unified score between two interacting proteins to ensure the coverage and reduce the effect of false positive interactions. For each protein that is produced in the functional network, the degree, betweenness, and closeness centrality scores were determined and the identified network key proteins, that is, the degrees that are beyond the average degree. For betweenness score, we considered scores greater than $n \times \pi_p$,and for the closeness score, we considered scores greater than $\frac{1}{\pi_p}$, n is the size of the network and π_p is the network shortest path average

Retrieval of Enriched Processes and Pathways of Targets

We then worked on identifying the enriched process of the targets that have been identified. Gene Ontology(GO) and protein GO Annotated (GOA) mapping which is provided by the UNiprotKB-GOA project was used to unveil enriched processes in which the set of protein targets were involved.

The hyper-geometric distribution was used to obtain the p-value by observing at least i proteins from a target gene set of size x, it contains reference dataset contains y annotated genes out of N genes. Given by

$$P[X \geq i] = 1 - \sum_{k=0}^{i-1} \frac{\binom{y}{k}\binom{N-y}{x-k}}{\binom{N}{k}} \tag{3.48}$$

X is a random variable, it is the number of genes annotated with the GO term under consideration in which a disease associated gene subset that is given.

Drug Repositioning

There have been many failures throughout the drug development pipeline, and billions of dollars have been invested in an average of about 9–12 years to bring a new drug to the market. Improving R&D productivity therefore remains the most important priority for pharmaceutical industry. In light of these challenges, drug repositioning, which involves the detection and development of new clinical indications for existing drugs, or for those that are in the development pipeline, has emerged as an increasingly important strategy for the new drug discovery (Li et al., 2016). Drug repositioning has the ability to substantially reduce the risks of development and the costs, and shorten the delay between drug discovery and availability. As part of the 84 drug products introduced to market in 2013, the new indications of existing drugs accounted for are 20%. Drug repositioning plays a key role in drug discovery and precision medicine criterion. Recently, drug repositioning is becoming strongly supported by governments, non-trading organizations and academic institutions. An example is both the United States (National Center for Advancing Translational Sciences) and the United Kingdom (Medical Research Council) who have launched large-scale funding programs in this area with the aim of extending molecules that already have undergone significant research and development by the pharmaceutical industry for more new indications (Li et al., 2016).

There are a number of drug repositioning strategies, these are, the computational repositioning strategies namely: genome, phenome, and computational repositioning approaches based on machine learning, network analysis, and text mining and semantic inference. The machine learning models can be used to leverage data to study underlying systems for prediction of new associations between drugs and diseases for the various data sources that support the exploring of repositioning approaches (Li et al., 2016). The network-based analysis strategy is another computational drug repositioning strategy which is widely used. The advances in high throughput technology and bioinformatics methods, molecular interactions in the biological system can all be modeled via networks. Past, studies suggested that drug target networks, drug-drug networks, drug disease network, protein interactions network and many other networks were useful in identifying characteristics of drug targets and provide new opportunities for drug repositioning (Li et al., 2016).

Li et al. also developed a bipartite drug target network method in order to identify potential new indications of an existing drug through its relation to similar drugs. In the model developed by Li et al., the drug pair similarity integrated the drug chemical structure similarity, common drug targets and their interactions. The team then built of the past success of the bipartite method by building a casual network which is a multi-layered pathway of gene, disease and drug target in order to identify new therapeutic uses of existing drugs. Other network related methods were introduced by Wu et al. and Jin et al., however, they will not be discussed here (Li et al., 2016). There is also the text mining and semantic inference approach which makes use of the biomedical and pharmaceutical knowledge which is available in literature or databases that contain a large amount of information on drugs and diseases and can be mined automatically and retrieved. This is as a result of the advancement in text mining research lately. Detecting new drug indications for existing drugs is highly possible by finding relevant knowledge via text mining. Computational drug repositioning is of great significance and can help improve human health through the discovery of new uses for existing drugs (J. Li et al., 2016). In this research, the network based model of computational drug repositioning is used to find new indications for influenza A.

Summary

In summary, this chapter provides a vivid description of the different datasets used in the analysis, and their sources. We also provide clearly the scoring scheme for sequence data, and the scores of the other data collected for this work. The unified score and it effectiveness is also presented in this chapter. The most important thing to note in this chapter is the scoring scheme for the sequence data, this is because it shows how the sequence data was scored specifically for this research work. Also, details of the different procedures used on the datasets are described in this chapter to obtain the results which are discussed in Chapter Four. We discuss the methods that describe how differentially abundant proteins are extracted, how candidate proteins are identified and procedure for enrichment analysis are all in this chapter. It is worth mentioning that, these are the main procedures through which potential target of IAV and potential drug targets were found.

CHAPTER FOUR

RESULTS AND DISCUSSION

To predict and analyse candidate proteins associated with IAV strains H1N1 and H3N2 infection at the systems level, we start by constructing human protein-protein functional interaction network. In this chapter, we used computational methods described in the previous chapter to integrate different protein-protein interaction datasets in order to generate unified networks and map different targets/markers identified into these networks to assess the contribution of each target/marker to the system using network centrality measures. Moreover, we checked whether these candidate proteins or targets are related by interacting in the map, belonging to the same cluster or being closely related based on GO processes annotating them. Since targets or markers, especially those associated with a specific strain (H1N1 or H3N2), may differ, but may be involved in the similar processes or work together in the same pathway, we also performed functional analyses of different targets, identifying enriched processes and pathways in which different markers are involved. We also predicted drugs which can be used for treating influenza A and diseases that are similar to H1N1, and H3N2

General topological properties of human functional networks

In this section, we present general features and topological properties of the human functional networks generated. It is worth mentioning that in different analysis, interactions with low confidence scores (between 0 and 0.3) were discarded unless they are predicted by at least two different sources. The human functional network is comprised of 2878644 functional interactions connecting 19170 proteins in the set of 20204 human reviewed proteins as read from the file retrieved from the UniProt database. Different predicted interactions were categorized in three classes, namely low, medium and high confidence interactions and, in total, 2931698, 1104684 and 1773958 functional interactions were identified for different categories, respectively.

Fitting degree and path-length distribution

Degree distribution: The result in Figure 3 shows that this human functional network satisfies scale-free topology properties, that is, the degree distribution of proteins approximates a power law function $\mathrm{P(k)} = k^{-\gamma}$, with the degree exponents $\gamma \sim 1.34506$. This degree exponent values was determined using the linear model:$\log\big(\mathrm{P}(k)\big) \sim -\gamma \ \log(k)$, linear inlog, with p -values $< 2e - 16$ for human and influenza A functional networks respectively, under the null hypothesis that $\gamma = 0$. This means that the observed data shows evidence that these power exponent values are significantly different from 0. This means that although some of the proteins would have many interacting partners, most of them would have few partners. The proteins that have many interacting partners are called “high degree” proteins or hubs or key proteins and probably ensure some basic chemical operations, such as energy transfer and redox reactions, essential for the survival of the organism (Newman, 2003b). The circle in Figure 3 represents the frequency $\mathrm{P}(k)$ of observing a protein interacting with k partners in a functional network. The solid line plots the power-law function approximating the connectivity distribution.

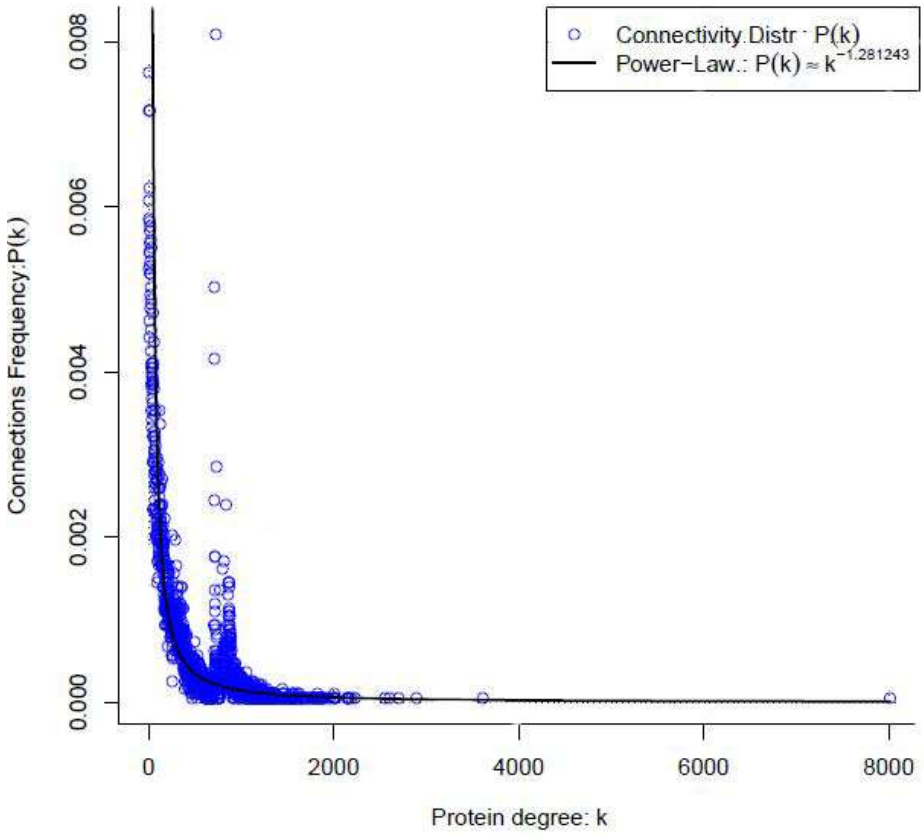

Figure 2: Degree distribution in the human functional network.

Path-length distribution: The average path lengths are approximately 3 (hops) for Influenza A and human functional networks, respectively, indicating that the spread of biological information in these systems is relatively fast. The two functional networks have a “small world” property, that is, the transmission of biological information from a given protein to others is achieved through only a few steps see Figure 3. This provides an idea about the network navigability, indicating how fast the information can be spread in the system independently of the number of proteins Mazandu &

Mulder (2011b). This indicates that the human system would be able to efficiently respond to the perturbations in the environment and to quickly exhibit a qualitative change of behaviour in response to these perturbations. The histogram plot represents the path-length distribution, that is, frequency of occurrence of shortest path of length ℓ, $\ell = 1,2,3 \ldots$ and the dashed line plot is the normal distribution approximating the path length distribution.

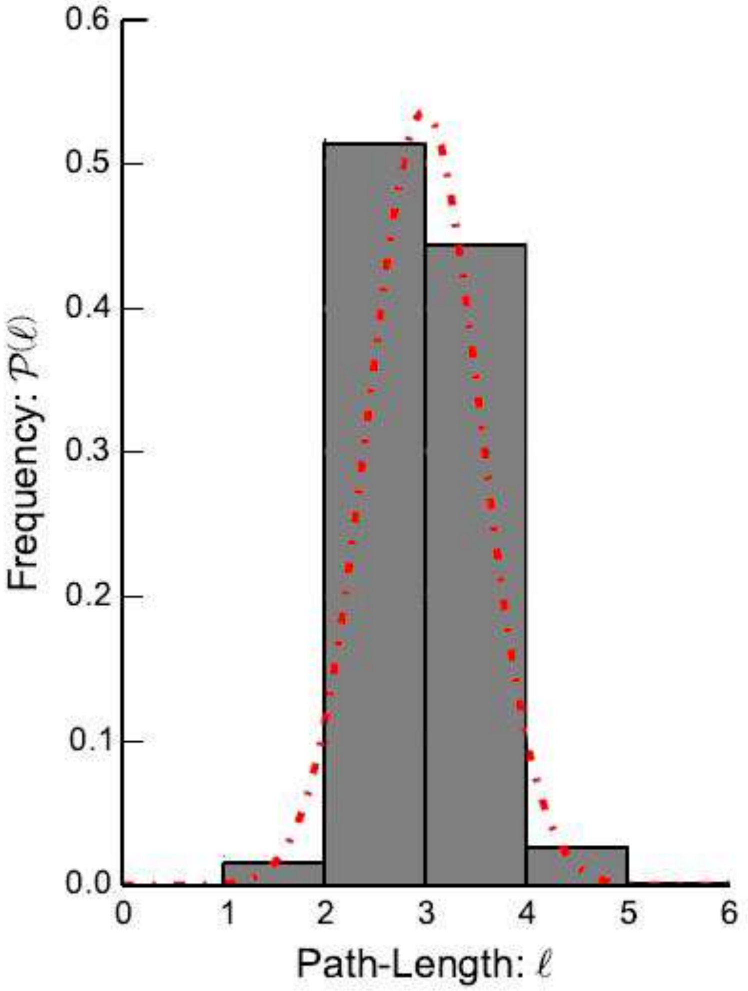

Figure 3: Path length distribution in the human functional network

Identification of disease-associated genes

Revealing the hierarchical community structure of the networks, identifying key proteins and mapping disease associated genes or proteins onto unified functional networks can provide a better understanding of the disease pathogenesis and increased potential to fully characterize the susceptible genes. In the human system, 2878644 interactions out of a total of 5810340 predicted functional interactions in the unified network with scores ranging from medium to high confidence level were used. The results showed that 5752 key proteins out of 19170 found in the human protein-protein functional network representing about 30% of interacting proteins in the human protein-protein functional interaction network. The clustering algorithm adopted from Blondel et al. (2008) is described in Chapter Three splits this human networks into 52 clusters or densely connected communities, the biggest or giant cluster containing 4385 proteins, of which, 678 are key proteins. Among the 52 clusters in the network, only 9 of the clusters contain key proteins and the distribution of these key proteins as shown in Table 3 and Table 4 for H1N1 and H3N2 respectively.

Table 3: H1N1 Clustering Results

CN	NP	KP	UP	UKP	DP	DKP	DIP	DIKP	TKP
0	4385	678	18	6	3	1	96	34	40
1	1333	408	1	1	1	19	19	11	13
2	4233	1178	7	2	17	5	264	113	117
3	2890	941	5	4	2	0	41	16	2
4	925	156	2	0	4	2	30	9	11
5	1941	66	0	0	2	0	13	1	1
6	987	563	1	1	2	2	40	33	3
7	1188	844	2	2	1	1	32	27	30
8	1288	928	0	0	1	1	7	4	5
Total	19170	5752	360	16	33	13	542	248	240

Table Key:

NP- number of proteins KP-key proteins UP-up key proteins

UKP- up key proteins DP- down proteins DKP-down key proteins

DIP-disease protein DIKP-disease key protein

TKP- total key proteins

Table 4: H3N2 Clustering Results

NP	KP	UP	UKP	DP	DKP	DIP	DIKP	TKP
0	4385	678	9	4	5	2	96	39
1	1333	408	1	1	1	1	19	13
2	4233	1178	5	1	17	5	264	115
3	2890	941	2	1	5	3	41	20
4	925	156	1	0	3	2	30	11
5	1941	66	2	0	1	0	13	1
6	978	553	1	1	4	4	40	38
7	1188	844	3	3	1	0	32	29
8	1288	9288	0	0	2	2	7	6
Total	19170	5752	24	11	39	19	542	272

Table Key:

NP- number of proteins KP-key proteins UP-up key proteins

UKP- up key proteins DP- down proteins DKP-down key proteins

DIP-disease protein DIKP-disease key protein

TKP- total key proteins

Filtering IAV disease-associated genes

Following the model of extraction of differentially (lower or higher) abundant genes described in Chapter Three, the two subsets of 69 and 64 out of 542 genes/proteins were extracted from H1N1 and H3N2 associated samples, respectively. More specifically, a total of 36 proteins are differentially up out of 95 up-regulated proteins whilst a total of 33 proteins are differentially down-regulated out of 105 down-abundant proteins in the case of H1N1. In the case of H3N2, a total of 24 proteins are differentially up out of 87 up-proteins, and 40 proteins are differentially down out of 113 up-abundant proteins. To assess the statistical significance of the subsets identified, a randomly generated 1000 independent samples of the same size as the identified set of differentially infected proteins and compute the aggregated score (ES) of each set generated. The normality test of ES of sample sets generated using the Shapiro-Wilk test and for both groups, the $p -$ value was 0.0 < 0.05 the significance level, indicating that these generated subsets are not normally distributed. Thus, we used the Wilcoxon test to check whether the identified set of differentially, infected proteins for each group can be expected by chance under the null hypothesis, comparing ES of sets of differentially abundant genes to those of randomly generated subsets. The corresponding p-value of 3.32586e-165 for both H1N1 and H3N2 associated samples, respectively, which are less than 0.05. Thus, the identified sets of differentially abundant genes for both H1N1 and H3N2 associated samples are more than expected by chance.

The differentially abundant proteins were extracted (up-regulated and down-regulated proteins). The p-value for the set was found to be 0.00 which implied that the null hypothesis of the Shapiro-Wilk test was to be rejected, suggesting that the set was not normally distributed. Hence, a Wilcoxon test performed gave the p-value of 3.32586e-165 and this implies that hence the null hypothesis was rejected. The differentially abundant proteins are shown on Table 5. This process was performed using H1N1 and H3N2 influenza A subtype data, the tables contain the list of all differentially abundant proteins for both H1N1 and H3N2. We provide in the table, the Uniprot ID of each differentially abundant protein, the gene name, and the protein description of each protein identified as up abundant or down abundant. There were proteins shared by the two sets (H1N1 and H3N2) for both up-regulated and down-regulated proteins. Some of these proteins are: Q00610 (Clathrin heavy chain), P01857 (Ig gamma-1 chain C region), and P02671 (Fibrinogen alpha chain). This suggested a certain level of similarity between the two influenza A subtypes. There were however some proteins that were H1N1 or H3N2 specific, these are P22314 (Ubiquitin-like modifier-activating enzyme 1) which is up-regulated and specific to H1N1, and Q13228 (Selenium binding protein 1) which is up-regulated and specific to H3N2. Up regulated proteins are represented by 1 and down regulated proteins are represented by 0.

Table 5: Differentially Abundant proteins for both H1N1 and H3N2

Protein ID	Gene Name	Protein Description	Status	Species
P05161	ISG15	Ubiquitin-like proteins ISG15	1, 1	H1N1, H3N2
P22314	UBAI	Ubiquitin-like modifier-activating enzyme 1	1,	H1N1,
P42224	STAT1	Signal transducer and activator transcription of 1-alpha/beta	1, 1	H1N1, H3N2
P05141	SLC25A5	ADP/ATP translocase2	1, 1	H1N1, H3N2
P52895	AKR1C2	Aldo-keto reductase family member C2	1,	H1N1
P08123	COL1A2	Collagen alpha-2(I) chain	1,	H1N1
P00352	ALDH1A1	Retinal dehydrogenase 1	1,	H1N1
P28838	LAP3	Cytosol aminopeptidase	1,1	H1N1, H3N2
P31946	YWHAB	14-3-3 protein beta/alpha	1	H1N1
P61978	HNRNPK	Heterogeneous nuclear ribonucleoprotein K	1, 1	H1N1, H3N2
P55072	VCP	Transitional endoplasmic reticulum	1,1	H1N1, H3N2
P40394	ADH7	Alcohol dehydrogenase class 4 mu/sigma chain	1, 1	H1N1, H3N2
P30838	ALDH3A1	Aldehyde dehydrogenase, dmeric NADP-preferring	1,	H1N1,H3N2

Table 5 continued

Protein ID	Gene Name	Protein Description	Status	Species
P52209	PGD	6-phosphogluconate dehydrose, decarboxylating	1, 1	H1N1, H3N2
Q00610	CLTC	Clathrin heavy chain	1, 1	H1N1, H3N2
P02452	COL1A1	Collagen alpha-1(I)	1	H1N1
Q14764	MVP	Major vault protein	1	H1N1
Q5VTE0	EEF1A1P5	Putative elongation factor 1-alpha-like 3	1, 1	H1N1, H3N2
O60437	PPL	Periplakin	1	H3N2
Q09666	AHNAK	Putative elongation factor 1-a+H15:L26lpha-li+J25ke 3	1	H3N2
Q9UBC9	SPRR3	Small proline-rich protein 3	1	H3N2
P20591	MXI	Interfeon-induced GTP-binding protein Mx1	1	H3N2
Q13228	SELENBP1	Selenium binding protein 1	1	H3N2
P22532	SPRR2D	Small proline-rich protein 2D	1	H3N2
P12814	ACTN1	Alpha-actinin-1	0	H3N2
P01023	A2M	Alpha-2-macroglobulin	0	H3N2

Table 5 continued

Protein ID	Gene Name	Protein Description	Status	Species
P01023	A2M	Alpha-2-macroglobulin	0	H3N2
P19338	NCL	Nucleolin	1, 1	H1N1,H3N2
Q71U36	TUBA1A	Tubulin alpha-1A	1, 1	H1N1, H3N2
P30044	PRDX5	Peroxiredoxin-5, mitochondrial	1, 1	H1N1, H3N2
P63104	YWHAZ	14-3-3 protein zeta/delta	1	H1N1
P00326	ADH1C	Alcohol dehydrogenase 1C	1	H1N1
Q04828	AKR1C1	Aldo-keto reductase family 1 member C1	1	H1N1
Q96KP4	CNDP2	Cytosolic non-specific dipeptidase	1	H1N1
P21980	TGM	Protein-glutamine gamma glutamyltansferase 2	1,1	H1N1,H3N2
O75874	IDHIPICD	Isocitrate dehydrogenase [NADP]cytoplasmic	1	H1N1
P62258	YWHAE	14-3-3 protein epsilon	1	H1N1

Table 5 continued

Protein ID	Gene Name	Protein Description	Status	Species
O60218	AKR1B10	Aldo-keto reductasen family 1 member	1	H1N1
P07195	LDHB	L-lactate dehydrogenase B Chain	1	H1N1
Q13938	CAPS	Calcyphosin	1	H1N1
Q01105	SET	Protein SET	0	H3N2
P02647	APOA1	Appolipoprotein A-1	0	H3N2
P07237	P4HB	Protein disulfide-isomerase	0	H1N1
Q08188	TGM3	Protein-glutamine gamma-glutamyltransferase	0	H1N1
P46940	IQGAP1	Ras GTPase-activating-like protein IQGAP1	0	H1N1
P02675	FGB	Fibrinogen beta chain	0	H1N1
P19013	KRT4	Keratin, type II cytoskeletal 4	0	H1N1
Q9UBG3	CRNN	Cornulin	0	H1N1
P29508	SERPINB3	Serpin B3	1	H1N1

Table 5 continued

Protein ID	Gene Name	Protein Description	Status	Species
Q9UBC9	SPRR3	Small proline-rich protein 3	0	H1N1
P09960	LTA4H	Leukotriene A-4 hydrolase	0	H1N1
P13646	KRT13	Keratin, type I cytoskeletal 13	0	H1N1
P9808	MUC5AC	Mucin-5AC	0	H1N1
Q5QNW6	HIST2H2BF	Histone H2B type2-F	0	H1N1
Q90BD6	RHCG	Ammonium transporter Rh type C	0	H1N1
Q9NR45	NANS	Sialic acid synthase	1	H1N1
P01871	IGHM	Ig mu chain C region	0	H3N2
P0C0S8	HIST1H2AG	Histone H2A type 1	0	H3N2
P06702	S100A9	Protein S100-A9	0,0	H1N1, H3N2
P62805	HIST1H4A	Histone H4	0, 0	H1N1, H3N2
P02452	COL1A1	Collagen alpha-1(I)	0	H3N2
P07384	CAPN1	Calpain-1 catalytic subunit	0	H3N2
P31146	CORO1A	Coronin-1A	0	H3N2

Table 5 continued

Protein ID	Gene Name	Protein Description	Status	Species
P01876	IGHA1	Ig alpha-1 chain C region	0	H3N2
P01877	IGHA2	Ig alpha-2 chain C region	0	H3N2
Q01518	CAP1	Adenylyl cyclase-associated protein 1	0	H3N2
Q13228	SELENBP1	Selenium-binding protein 1	1	H1N1
P40925	MDH1	Malate dehydrogenase, cytoplasmic	1, 1	H1N1, H3N2
P02675	FGB	Fibrinogen beta chain	0	H3N2
P14923	JUP	Junction plakoglobin	0, 0	H1N1, H3N2
P12429	ANXA3	Annexin A3	0, 0	H1N1, H3N2
P0C0L4	C4A	Complement C4-A	0	H3N2
P59665	DEFA1	Neutrophil defensin 1	0, 0	H1N1, H3N2
P00338	LDHA	L-lactate dehydrogenase A chain	0	H3N2
P08133	ANXA6	Annexin A6	0, 0	H1N1, H3N2

Table 5 continued

Protein ID	Gene Name	Protein Description	Status	Species
P27105	STOM	Erythrocyte band 7 integral membrane protein	0, 0	H1N1, H3N2
P02671	FGA	Fibrinogen alpha chain	0, 0	H1N1, H3N2
P02750	LRG1	Leucine-rich alpha-2-glycoprotein	0, 0	H1N1, H3N2
P09960	LTA4H	Leukotriene A-4 hydrolase	0	H3N2
P02679	FGG	Fibrinogen gamma chain	0, 0	H1N1, H3N2
P49913	CAMP	Cathelicidin antimicrobial peptide	0, 0	H1N1, H3N2
P00450	CP	Ceruloplasmin	0	H3N2
P14780	MMP9	Matrix metalioproteinase-9	0, 0	H1N1, H3N2
P52566	RHGDIB	Rho GDP-dissociation inhibitor 2	0	H3N2
P11215	ITGAM	Integrin alpha-M	0, 0	H1N1, H3N2
P00738	HP	Haptoglobin	0	H3N2
P01857	IGHG1	Ig gamma-1 chain C region	0, 0	H1N1, H3N2
P00761	-	Trypsin	0	H3N2

Table 5 continued

Protein ID	Gene Name	Protein Description	Status	Species
P04083	ANXA1	Annexin A1	0	H3N2
P00491	PNP	Purine nucleoside phosphorylase	0	H3N2
Q01518	CAP1	Adenylyl cyclase-associated protein 1	0	H3N2
P31146	CORO1A	Coronin-1A	0	H3N2

Enrichment and functional closeness analyses of differentially abundant proteins

For each GO term process, the p-value was calculated using the hyper-geometric distribution and this was based on its frequencies of occurrence in the reference dataset (human proteome) and target set, which was made up of identified protein targets, and adjusted using the Bonferroni multiple testing correction. For relationships between GO terms in the GO structure, the concept of the GO term semantic similarity score was used, specifically the GO-universal metric to compute the frequency of occurrence of the target-associated process in a set of proteins based on the semantic similarity degree.

For enriched processes in H1N1 found on Table 6, we came across the process GO:0043277 (apoptotic cell clearance) which is the recognition and removal of an apoptotic cell by a neighboring cell or by a phagocyte (http://amigo1.geneontology.org/cgibin/amigo/term_details?term=GO:0043277) and GO:0071395 (cellular response to jasmonic acid stimulus) which is any process that results in a change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a jasmonic acid stimulus (http://amigo1.geneontology.org/cgi-bin/amigo/term_details?term=GO:0071395). For enriched processes in H3N2, the process GO:0006958 (complement activation, classical pathway) highly contributes to classical pathway of the complement cascade which allows for the direct killing of microbes, disposal of immune complexes and the regulation of the other immune processes (http://amigo1.geneontology.org/cgi-bin/amigo/term_details?term=GO:0006958). Complement is a system of circulating enzymes and is part of the body's response to illness or injury, it plays a very important role which is in the host's defense against infectious agents and in the inflammatory process. Classical pathway of complement activation however has often been regarded as the major enforcer for antibody action. (Brown and Gilliland, 2002), show in their work, the vital role of classical pathway in the innate immunity of *Streptococcus pneumonia* infection in mice. The process GO:0051607 (defense response to virus) is highly important and contributes in the process as a reaction triggered in response to the presence of a virus that act to protect the cell or organism (http://amigo1.geneontology.org/cgibin/amigo/term_details?term=GO:0051607). The enriched processes for H3N2 are shown on Table 7.

Table 6: Enriched Processes in Target Set for H1N1

GO ID	GO Name	GO Level	P-Value	Bonferroni Correction
GO:0001878	response to yeast	6	1.41129E-05	0.00635
GO:0043152	induction of bacterial agglutination	8	2.85876E-08	1.2864E-05
GO:2000352	negative regulation of endothelial cell apoptotic process	8	2.27150E-05	0.01022
GO:0072377	blood coagulation, common pathway	7	2.85938E-08	1.2867E-05
GO:0045907	positive regulation of vasoconstriction	7	5.76781E-05	0.02596
GO:0043206	extracellular fibril organization	6	1.45452E-08	6.5453E-06
GO:0006069	ethanol oxidation	7	3.94877E-06	0.00178
GO:0034116	positive regulation of heterotypic cell-cell adhesion	5	2.03096E-05	0.00913
GO:0044597	daunorubicin metabolic process	7	3.99372E-07	0.00018
GO:0044598	doxorubicin metabolic process	6	3.99372E-07	0.00018
GO:0031424	keratinization	8	3.44408E-05	0.0155

Table 6 continued

GO ID	GO Name	GO Level	P-Value	Bonferroni Correction
GO:0043277	apoptotic cell clearance	7	3.30081E-05	0.01485
GO:1902309	negative regulation of peptidyl-serine dephosphorylation	9	0.0001	0.04724
GO:0042730	fibrinolysis	8	4.49378E-05	0.02022
GO:0050829	defense response to Gram-negative bacterium	7	3.8545E-12	1.7345E-09
GO:0050830	defense response to Gram-positive bacterium	7	4.1286E-10	1.8579E-07

Table 7: Enriched Processes in Target Set for H3N2

GO ID	GO Name	Go Level	P-Value	Bonferroni Correction
GO:0006958	complement activation classical pathway	8	2.80171e-05	0.01042
GO:0022617	extracellular matrix disassembly	5	2.77671e-07	0.00010
GO:0071353	cellular response to interleukin-4	6	0.00011	0.04450
GO:0001878	response to yeast	6	1.07253e-05	0.00399
GO:0043152	induction of bacterial agglutination	8	4.34419e-11	1.61604e-08
GO:0072377	blood coagulation, common pathway	7	2.168220e-08	8.06578e-06
GO:0050853	B cell receptor signaling pathway	9	7.77836e-06	0.00289
GO:0050871	positive regulation of B cell activation	7	2.01938e-05	0.00751
GO:0043277	apoptotic cell clearance	7	2.30170e-05	0.00856
GO:0042730	fibrinolysis	8	3.41989e-05	0.01272
GO:0050829	defense response to Gram-negative bacterium	7	0.0	0.0
GO:0007597	blood coagulation, intrinsic pathway	7	6.61354e-08	2.46024e-05

Table 7 continued

GO ID	GO Name	GO Level	P-Value	Bonferroni Correction
GO:0060267	positive regulation of respiratory burst	4	8.74429e-05	0.03253
GO:0050830	defense response to Gram-positive bacterium	7	8.78975e-12	3.26979e-09
GO:0051607	defense response to virus	6	2.87036e-05	0.01068

Systems level based identification of protein targets

Differentially abundant proteins in samples from influenza A virus (IAV) positive patients versus IAV negative patients were analysed at the systems level to identify the final set of targets using several criteria, including statistical and biological relevance, this was done for both H1N1 and H3N2. The set of differentially abundant proteins were then filtered based on their abundance levels and checking whether they have been previously identified. Mapping these proteins to the generated human protein-protein functional network to elucidated their roles in the system based on the centrality measures that have been discussed in Chapter Three. The clusters and key proteins displayed in this section were all found based on methods described in Chapter Three as well.

A total of 2878644 interactions were found between human and influenza A. Also, 542 influenza disease-associated proteins were found from DisGeNet. The number of key proteins shared between differentially (up-abundant and down-abundant) proteins are shown on Table 8. The disease key proteins which are up-abundant and down abundant were also found for H1N1 and H3N2 out of a total of 248 disease key proteins. Furthermore, the protein BPIFA1 or SPLUNC1 which was found in Leeming et al. (2015) and Liu et al. (2013) were also identified in this work as a down-regulated protein for both H1N1 and H3N2 influenza A subtypes. The protein BPIFA1 being down regulated implies that its contribution to processes that are necessary for the system to fight against infections is very minimal or it contributes nothing at all. The work of this particular protein however is to inhibit other signal processes that affect the human, due to its inability to function properly, it allows that pathogen to have its way. The diseased key proteins are shown in Table 9.

Table 8: Shared key proteins between differentially abundant proteins and DisGeNet

Proteins ID	Protein Name	Gene Name	Status	Species
O60218	Aldo-keto reductase family 1 member B10	AKR1B10	1	H1N1
P05161	Ubiquitin-like protein ISG15	ISG15	1,1	H1N1, H3N2

Table 8 continued

Proteins ID	Protein Name	Gene Name	Status	Species
P05164	Myeloperoxidase	MPO	0,0	H1N1, H3N2
P42224	Signal transducer and activator of transcription 1-alpha/beta	STAT1	1,1	H1N1, H3N2
P05109	Protein S100-A8	S100A8	0,0	H1N1, H3N2
Q9NP55	BPI fold-containing family A member	BPIFA1	0,0	H1N1, H3N2
P14780	Matrix metalloproteinase -9	MMP9	0,0	H1N1, H3N2
P00738	Haptoglobin	HP	0	H3N2
P20591	Interferon-induced GTP-binding protein Mx1	MX1	1	H3N2

Table 9: Disease key proteins and differentially abundant proteins

Proteins ID	Protein Name	Gene Name	Status	Species
P05161	Ubiquitin-like protein ISG15	ISG15	1,1	H1N1, H3N2
P42224	Signal transducer and activator of transcription 1-alpha/beta	STAT1	1,1	H1N1, H3N2
P05109	Protein S100-A8	S100A8	0,0	H1N1, H3N2
Q9NP55	BPI fold-containing family A member 1	BPIFA1	0,0	H1N1, H3N2
P05164	Myeloperoxidase	MPO	0,0	H1N1, H3N2
P14780	Matrix metalloproteinase-9	MMP9	0,0	H1N1, H3N2
P00738	Haptoglobin	HP	0	H3N2
P20591	Interferon-induced GTP-binding protein Mx1	MX1	1	H3N2

The details of the differentially abundant proteins and the clusters they are found in Table 10.

Table 10: Clusters with up and down abundance key proteins for H1N1 and H3N2

Cluster Number	Proteins ID	Protein Name	Status	Species
0	P05161	Ubiquitin-like protein ISG15	1	H1N1, H3N2
0	P22314	Ubiquitin-like modifier-activating enzyme1	1	H1N1
0	P6197	Heterogeneous nuclear ribonucleoprotein K	1	H1N1, H3N2
0	P40925	Malate dehydrogenase, cytoplasmic	1	H1N1, H3N2
0	P19338	Nucleolin	1	H1N1, H3N2
0	P07195	L-lactate dehydrogenase B chain	1	H1N1
0	P07237	Protein disulfide-isomerase	0	H1N1
0	P00338	L-lactate dehydrogenase A chain	0	H3N2
0	P07384	Calpain-1 catalytic subunit	0	H3N2
1	Q00610	Clathrin heavy chain 1	1	H1N1, H3N2
1	P14923	Junction plakoglobin	0	H1N1, H3N2
2	P42224	Signal transducer and activator of transcription 1-alpha/beta	1	H1N1, H3N2
2	P02452	Collagen alpha-1(I) chain	1	H1N1, H3N2
2	P04083	Annexin A1	0	H1N1

Table 10 continued

Cluster Number	Proteins ID	Protein Name	Status	Species
2	P11215	Integrin alpha-M	0	H1N1, H3N2
2	P14780	Matrix metalloproteinase -9	0	H1N1, H3N2
2	P02647	Apolipoprotein A-I	0	H1N1
2	P05109	Protein S100-A8	0	H1N1, H3N2
2	P00738	Haptoglobin	0	H3N2
3	P63104	14-3-3 protein zeta/delta	1	H1N1
3	P62258	14-3-3 protein epsilon	1	H1N1
3	P31946	14-3-3 protein beta/alpha	1	H1N1
3	Q71U36	Tubulin alpha-1A chain	1	H1N1, H3N2
3	P12814	Alpha-actinin-1	0	H3N2
3	P13796	Plastin-2	0	H3N2

Disease-associated gene annotation enrichment analysis

In order to gain insight into the biological processes of the proteins involved in human-influenza A interaction, we performed an enrichment analysis using GO annotated mapping provided by the UNiprotKB-GOA. For each protein, we performed a GO biological process enrichment analysis using the human proteins from the constructed network. We used the Bonferroni p-value which is also the corrected value for multiple testing and selected the GO terms enriched in our key protein list by fixing a p-value less than 0.05. A total of 225 biological processes were found for H1N1 and 233 for H3N2 in the enrichment analysis. A section of the GO enriched processes are listed in Table 11 for H1N1 and Table 12 for H3N2. The biological processes include those which are particularly relevant to the intracellular environment of IAV in the host, such as cellular defense response.

Table 11: Analysis Result: GO Enriched Processes in Target Set for H1N1

GO ID	GO Name	Go Level	P-Value	Bonferroni Correction
GO:0022614	membrane to membrane docking	4	9.03661e-06	0.01304
GO:0045356	positive regulation of interferon-alpha biosynthetic process	7	1.74087e-08	2.51382e-05
GO:0071360	cellular response to exogenous dsRNA	7	5.956294e-06	0.00860
GO:0050852	T cell receptor signaling pathway	9	2.03425e-10	2.93746e-07
GO:0007218	neuropeptide signaling pathway	6	0.0	0.0
GO:0045740	positive regulation of DNA replication	7	2.74857e-06	0.00397
GO:0097284	hepatocyte apoptotic process	7	0.0	0.0
GO:0045814	negative regulation of gene expression, epigenetic	6	2.22632e-05	0.03215
GO:2000553	positive regulation of T-helper 2 cell cytokine production	9	3.31636e-06	0.00479
GO:0035690	cellular response to drug	4	1.76943e-05	0.02555
GO:0071230	cellular response to amino acid stimulus	6	4.21451e-07	0.00061
GO:0035711	T-helper 1 cell activation	12	1.41027e-11	2.03643e-08

Table 12: Analysis Result: GO Enriched Processes in Target Set for H3N2

GO ID	GO Name	Go Level	P-Value	Bonferroni Correction
GO:0048873	homeostasis of number of cells within a tissue	6	3.44493e-05	0.04874
GO:0045356	positive regulation of interferon-alpha biosynthetic process	7	1.71684e-08	2.42932e-05
GO:0080184	response to phenylpropanoid	5	2.91687e-11	4.12737e-08
GO:0050852	T cell receptor signaling pathway	9	2.11397e-10	2.99126e-07
GO:0097252	oligodendrocyte apoptotic process	7	0.0	0.0
GO:0045740	positive regulation of DNA replication	7	2.69111e-06	0.00381
GO:0045814	negative regulation of gene expression, epigenetic	6	2.15763e-05	0.03053
GO:0071230	cellular response to amino acid stimulus	6	4.11159e-07	0.00058
GO:0000060	protein import into nucleus, translocation	10	6.03673e-07	0.00085
GO:0035711	T-helper 1 cell activation	12	0.0	0.0
GO:0071442	positive regulation of histone H3-K14 acetylation	11	3.26488e-05	0.04619
GO:0060301	positive regulation of cytokine activity	7	8.93782e-06	0.01265

Viruses attack intracellular signaling and cytoskeletal pathways to change the responses of the host in a way that benefits them. Influenza A virus (IAV) interferes with pathways such as NF-kappa B (NF-kB) signaling pathway and MAPK signaling pathway which can lead to the prevention of NF-kB dependent transcription (Rapanoel et al. 2013). The Kyoto Encyclopedia of Genes and Genomes (KEGG) was used to find the pathways in which the protein in the human interaction network belong. For each of the key proteins, we obtained a list of pathways relevant to the key protein list. Not all proteins belong to a pathway and a protein may belong to several pathways. This was done for H1N1 and H3N2. Among the pathways found was the pathway for influenza A which is an interesting discovery. Other notable pathways which are used as a benchmark in this work are, MAPK signaling pathway, Apoptosis, NF-kappaB signaling pathway, Chemokine signaling pathway. The following subsections give a better description of some of the pathways mentioned here, Table 13 and Table 14 hold the results of the pathway analysis.

Table 13: Pathway Results for H1N1

KEGG Pathway ID	KEGG Pathway Name	P-Value	Adjusted p-values
hsa04611	Platelet activation	7.20076e-08	1.57696e-05
hsa05100	Bacterial invasion of epithelial cells	4.86503e-05	0.01065
hsa04612	Antigen processing and presentation	0.0	0.0
hsa05223	Non - small cell lung cancer	2.28721e-10	5.00901e-08
hsa05222	Small cell lung cancer	3.91661e-06	0.00085
hsa04514	Cell adhesion molecules (CAMs)	1.03364e-09	2.26368e-07
hsa05220	Chronic myeloid leukemia	1.23185e-08	2.69775e-06
hsa04510	Focal adhesion	1.40409e-08	3.07497e-06
hsa04921	Oxytocin signaling pathway	4.40910e-07	9.65594e-05
hsa04920	Adipocytokine signaling pathway	1.60523e-05	0.00351
hsa04660	T cell receptor signaling pathway	0.0	0.0
hsa04662	B cell receptor signaling pathway	1.81939e-06	0.00039
hsa04664	Fc epsilon RI signaling pathway	0.0	0.0
hsa04668	TNF signaling pathway	0.0	0.0
hsa05133	Pertussis	0.0	0.0
hsa04010	MAPK signaling pathway	0.0	0.0
hsa04210	Apoptosis	0.0	0.0
hsa04064	NF-kappaB signaling pathway	2.31568e-10	5.07134e-08
hsa04062	Chemokine signalling pathway	5.823451e-07	0.00012
hsa05164	Influenza A	0.0	0.0

Table 14**:** Pathway Results for H3N2

KEGG Pathway ID	KEGG Pathway Name	P-Value	Adjusted p-values
hsa04612	Antigen processing and presentation	0.0	0.0
hsa04750	Inflammatory mediator regulation of TRP channels	0.0	0.0
hsa05222	Small cell lung cancer	2.91987e-06	0.00064
hsa04514	Cell adhesion molecules (CAMs)	6.23205e-10	1.36482e-07
hsa04921	Oxytocin signaling pathway	2.95400e-07	6.46927e-05
hsa04920	Adipocytokine signaling pathway	1.24667e-05	0.00273
hsa04660	T cell receptor signaling pathway	0.0	0.0
hsa04662	B cell receptor signaling pathway	1.36300e-06	0.00029
hsa04664	Fc epsilon RI signaling pathway	0.0	0.0
hsa04668	TNF signaling pathway	0.0	0.0
hsa04723	Retrograde endocannabinoid signaling	1.10452e-06	0.00024
hsa04010	MAPK signaling pathway	0.0	0.0
hsa04210	Apoptosis	0.0	0.0
hsa04064	NF-kappaB signaling pathway	2.31568e-10	5.07135e-08
hsa04062	Chemokine signalling pathway	5.82345e-07	0.00013
hsa05164	Influenza A	0.0	0.0

Nuclear Factor-kappa B (NF-kappa B) Pathway

Nuclear factor-kappa B (NF-kappa B) is the generic name of a family of transcription factors that function as dimers and regulate genes involved in immunity, inflammation and cell survival. There are several pathways leading to NF-kappa B-activation. This pathway relies on IKK- mediated IkappaB-alpha phosphorylation and this to its degradation, which allows the NF-kappa B dimer to enter the nucleus and activate gene transcription.

Influenza A Pathway

As already discussed, influenza A is a contagious respiratory disease caused by influenza virus infection. Novel strains that cause pandemics arise from avian influenza virus by genetic reassortment among influenza viruses and two surface glycoproteins HA and NA form the basis of serologically distinct virus types. The innate immune system recognizes invaded virus through multiple mechanisms. Viral non-structural NS1 protein is a multifunctional virulence factor that interfere IFN-mediated antiviral response. It inhibits IFN production by blocking activation of transcription factors such as NF-kappa B, IRF3 and AP1, and further inhibits the activation of IFN-induced antiviral genes. PB1-F2 protein induces apoptosis of infected cells, which results in life-threatening bronchiolitis.

Chemokine Signaling Pathway

Chemokines are small chemoattractant peptides that provide directional cues for the cell trafficking and thus are vital for protective host response. In addition, chemokines regulate plethora of biological processes of hematopoietic cells to lead cellular activation, differentiation and survival.

Induction of nitric oxide and production of reactive oxygen species are also regulated by chemokine signal via calcium mobilization and diacylglycerol production.

MAPK Signaling Pathway

The mitogen-activated protein kinase (MAPK) cascade is a highly conserved module and is involved in various cellular functions, including cell proliferation, differentiation and migration. Mammals express at least four distinctly regulated groups of MAPKs, that is, extracellular signal-related kinases, Jun amino-terminal kinases, p38 proteins and ERK5, that are activated by specific MAPKKs. Each MAPKK, however, can be activated by more than one MAPKKK, and this increases the complexity and diversity of MAPK signaling.

Apoptosis Pathway

Apoptosis is a genetically programmed process for the elimination of damaged or redundant cells by activation of caspases (aspartate-specific cysteine proteases). Apoptosis is controlled by numerous interrelating processes. The `extrinsic' pathway involves stimulation of members of the tumor necrosis factor (TNF) receptor subfamily. The `intrinsic' pathway is activated mainly by non-receptor stimuli, such as DNA damage, ER stress, metabolic stress, UV radiation or growth-factor deprivation. The central event in the 'intrinsic' pathway is the mitochondrial outer membrane permeabilization (MOMP), which leads to the release of cytochrome c. These two pathways converge at the level of effector caspases, such as caspase-3 and caspase-7. The third major pathway is initiated by the constituents of cytotoxic granules released by CTLs (cytotoxic T-cells) and NK (natural killer) cells.

Functional relationships between identified disease-associated genes and other diseases

Understanding the relationship between diseases based on undiscovered biological mechanisms is one of the challenges of modern day biology and medicine. To explore disease-disease associations, we use

systems level biological data and this is expected to improve current knowledge of disease relationships of influenza A, which may lead to further improvements in disease diagnosis, prognosis and treatment. One of the commonly used biological data is disease-gene association.

Using the GO-universal metric-based BMA model discussed in Chapter Three. We compute semantic similarity scores between influenza A and 12483 diseases for H1N1 and H3N2 out of 12514 found in the list of disease-gene associations from the DisGeNet database after discarding genes without GO annotations (processes) in the human proteome. These scores were computed by considering (i) a set of disease-associated processes as retrieved from the human proteome using the semantic similarity score, (ii) a set of filtered disease-associated processes and (iii) set of enriched disease-associated processes. A disease is said to be similar to influenza A if the semantic similarity score between their associated enriched processes is high, this indicates that their sets of candidate disease proteins are functionally very close. A threshold value, τ_s, defined below in terms of third quartile (Q_3) and interquartile (IQR), is generally used to identify approved drugs that may be mapped to a given set of candidate. This threshold is given by

$$\tau_s = Q_3 + \varepsilon \times IQR \qquad (4.1)$$

where the interquartile range $IQR = Q_3 - Q_1$ with Q_1 as the first quartile and ε is a tuning parameter ranging between 0 and 1 that is, $0 \leq \varepsilon \leq 1.5$. For this research work, we set the threshold to be over $Q_3 + 0.75 \times IQR,$ where $\varepsilon = 0.75$. The threshold values obtained for H1N1 was 0.46325 and that of H3N2 is 0.45818, and disease similarity results are shown on Table 15 and Table 16 for H1N1 and H3N2 respectively. Figures 4 and 5 are the graphical representation of the disease similarity score of the candidate proteins found.

Table 15: Disease Similarity for H1N1

Disease-ID	Disease-Name	Enriched Similarity Scores	Non-Redundant Similarity Scores	Redundant Similarity Scores
C0021400	Influenza, Human	0.67545	0.91558	0.94996
C0014038	Encephalitis	0.67523	0.76149	0.83468
C0011311	Dengue	0.67447	0.77154	0.84271
C0011603	Dermatitis	0.6727	0.77051	0.84262
C1290886	Chronic inflammatory disorder	0.67116	0.71505	0.79496
C0006272	Bronchiolitis Obliterans	0.67048	0.72571	0.799
C0022408	Joint Diseases	0.66958	0.77326	0.84862
C0039103	Synovitis	0.66867	0.76581	0.83773
C0042384	Vasculitis	0.66733	0.74138	0.82346
C0035235	Respiratory Syncytial Virus Infections	0.66732	0.76906	0.84076
C0027697	Nephritis	0.66732	0.73878	0.81959
C0524909	Hepatitis B, Chronic	0.66471	0.77626	0.85144
C0993582	Arthritis, Experimental	0.6643	0.74869	0.82923
C1096184	West Nile viral infection	0.66407	0.72115	0.81016
C0026691	Mucocutaneous Lymph Node Syndrome	0.66407	0.77124	0.84577
C0031099	Periodontitis	0.66329	0.76498	0.84325
C0027430	Nasal Polyps	0.66297	0.7868	0.85424
C0020517	Hypersensitivity	0.66295	0.73915	0.81472
C0151317	Chronic infectious disease	0.66229	0.76885	0.8445

Table 16: Disease Similarity for H3N2

Disease-ID	Disease-Name	Enriched Similarity Scores	Non-Redundant Similarity Scores	Redundant Similarity Scores
C0021400	Influenza, Human	0.68084	0.91848	0.95122
C0011311	Dengue	0.68069	0.77346	0.84309
C0011603	Dermatitis	0.67757	0.77016	0.84158
C0014038	Encephalitis	0.67625	0.75761	0.83144
C0039103	Synovitis	0.67371	0.76684	0.83808
C0031099	Periodontitis	0.67269	0.76526	0.84396
C1290886	Chronic inflammatory disorder	0.67257	0.71856	0.79659
C0022408	Joint Diseases	0.6718	0.77622	0.85011
C0524909	Hepatitis B, Chronic	0.67119	0.76859	0.8468
C0019100	Dengue Hemorrhagic Fever	0.67088	0.73412	0.81267
C0027697	Nephritis	0.67083	0.73865	0.81838
C0006272	Bronchiolitis Obliterans	0.67035	0.72487	0.79853
C0042384	Vasculitis	0.66971	0.74201	0.82394
C0035235	Respiratory Syncytial Virus Infections	0.66941	0.77188	0.84198
C0017661	Glomerulonephritis, IGA	0.66841	0.75959	0.83861
C1096184	West Nile viral infection	0.66775	0.72407	0.81141
C0017658	Glomerulonephritis	0.66696	0.76555	0.84503
C0024143	Lupus Nephritis	0.66688	0.75681	0.83407

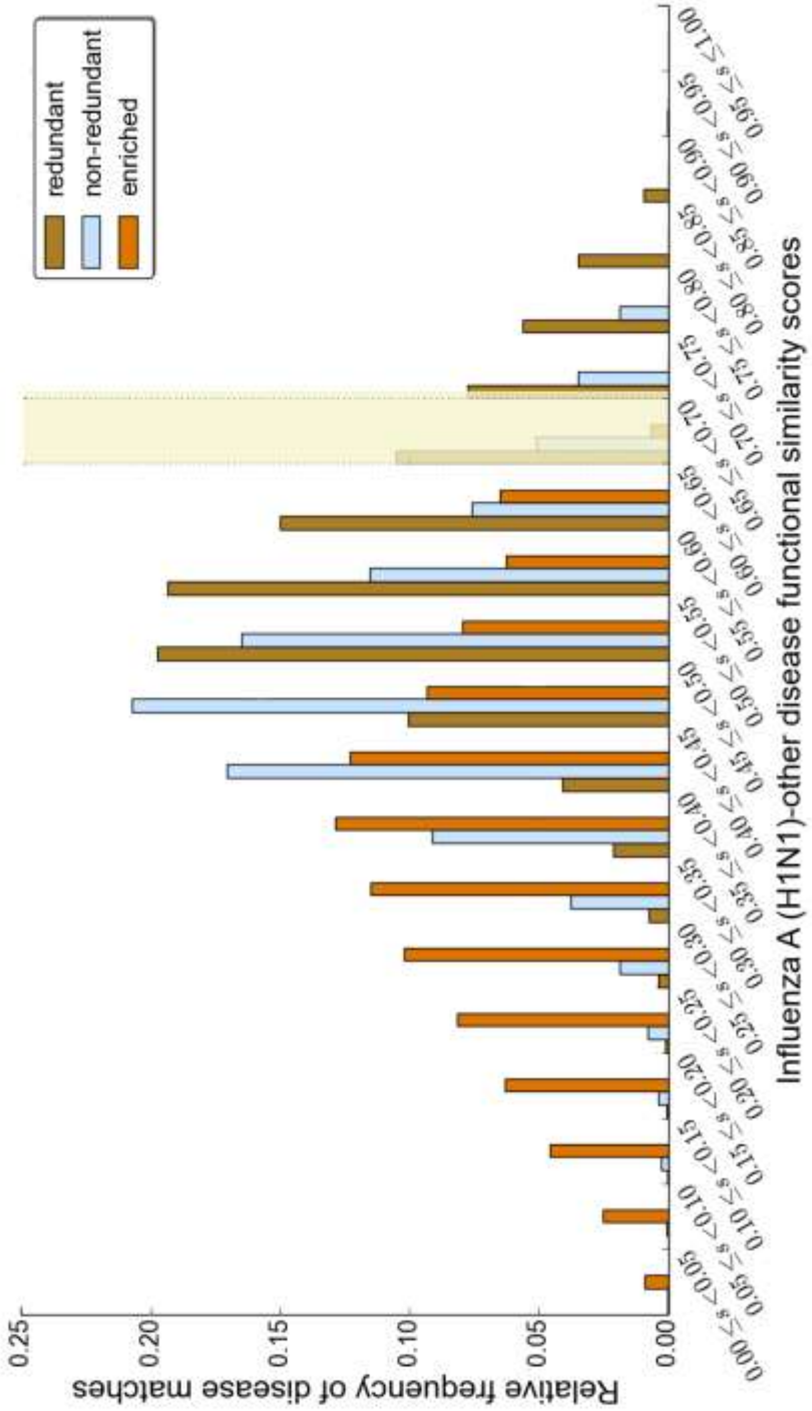

Figure 4: Disease Similarity Results for H1N1

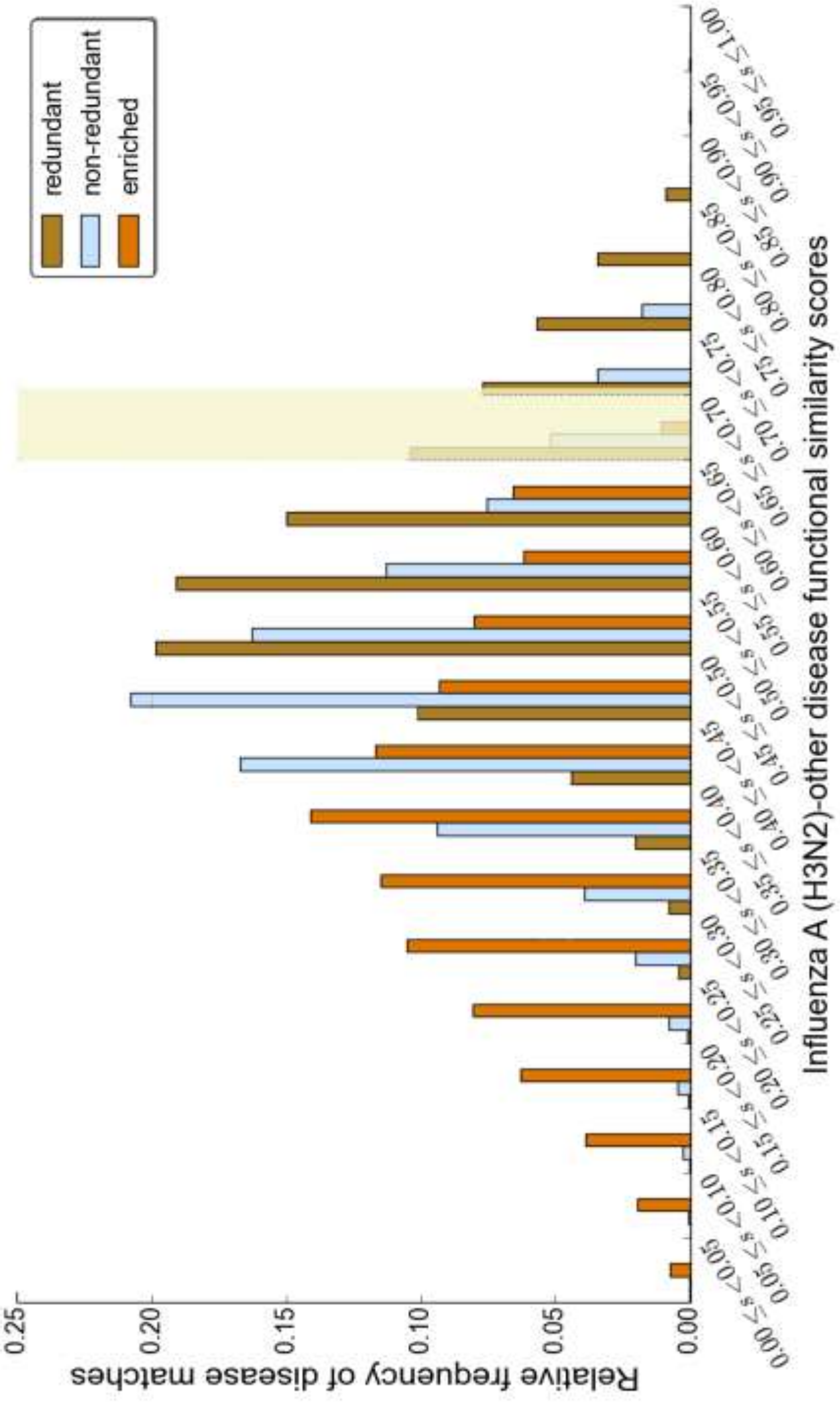

Figure 5: Disease Similarity Results for H3N2

Disease-associated genes and drug targets

One of the main reasons for conducting this research is to find putative drug targets in order to design new drugs from approved drug sources for treating influenza A. The drug repositioning was based on similarity scores, the aim was to find the high similarity between targets and drug targets. The disease based and drug based (network computational) method was implemented. This is important in the case of influenza A because of the existence of drug resistant influenza A. An example is an oseltamivir resistant influenza A type which was discovered by Meijer et al. (2009). In predicting potential drugs, each approved drug was mapped to a set of biological processes in which its proteins are involved, semantic similarity scores between the drug-associated processes and influenza A (H1N1 and H3N2) associated processes was also computed in order to associate influenza A with its potential drugs. A list of 1638 potential targets was predicted computationally for H1N1 and H3N2 using the human network, this was based on the similarity scores. The list was used to manually curate 71 and 82 drug hits for H1N1 and H3N2 respectively. From identified drugs sharing some similarity in terms of processes, we extracted and ordered those that are over $0.75 \times \mathrm{IQR}$, that is, showed a high similarity in terms of scores obtained, with $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$ is the interquartile range where Q1 and Q3 are respectively, the 1st (lower) and 3rd (upper) quartile. A portion of the drug similarity scores are shown for H1N1 and H3N2 in Table 15, and Table 16, here we see the similarity scores for the redundant (RSS), non-redundant (NRSS) and the enriched proteins (ESS). A drug known as Pranlukast was found as an indication drug. It antagonizes or reduces bronchospasm caused, principally in asthmatics, by an allergic reaction to accidentally or inadvertently encountered allergens. It is found in the following categories: Anti-Asthmatic Agents, Respiratory System, Drugs for Obstructive Airway Diseases, Leukotriene Receptor Antagonists and others, hence it can be used to treat a respiratory disease such as H1N1 subtype of influenza A. Figures 6, and Figure 7 provides a graphical representation of similarity scores and cut off for each H1N1 and H3N2

Table 17: Drug Similarity for H1N1

Drug-ID	Drug-Name	ESS	NRSS	RSS
DB00608	Chloroquine	0.56008	0.57865	0.6592
DB01296	Glucosamine	0.55798	0.61222	0.69694
DB05676	Apremilast	0.55156	0.60003	0.6924
DB01411	Pranlukast	0.55141	0.60778	0.68621
DB00033	Interferon gamma-1b	0.54745	0.52099	0.54834
DB00098	Anti-thymocyte Globulin (Rabbit)	0.54568	0.56633	0.65287
DB06681	Belatacept	0.54358	0.51685	0.60194
DB01281	Abatacept	0.54358	0.51685	0.60194
DB00005	Etanercept	0.54223	0.62547	0.70936
DB01041	Thalidomide	0.53912	0.60482	0.69174
DB01427	Amrinone	0.538	0.57487	0.6574
DB09052	Blinatumomab	0.53742	0.52392	0.55728
DB01017	Minocycline	0.53741	0.61398	0.71028
DB08904	Certolizumab pegol	0.53706	0.5925	0.66971
DB06674	golimumab	0.53706	0.5925	0.66971
DB00065	Infliximab	0.53706	0.5925	0.66971
DB00051	Adalimumab	0.53049	0.60275	0.6797
DB00852	Pseudoephedrine	0.53019	0.60884	0.69843
DB01169	Arsenic trioxide	0.52559	0.59667	0.69496
DB06168	Canakinumab	0.51552	0.57613	0.67037

Table 18: Drug Similarity for H3N2

Drug-ID	Drug-Name	ESS	NRSS	RSS
DB00608	Chloroquine	0.56212	0.57612	0.65836
DB01296	Glucosamine	0.56203	0.61411	0.69863
DB01411	Pranlukast	0.55823	0.60851	0.68729
DB00098	Anti-thymocyte Globulin (Rabbit)	0.55756	0.56721	0.6537
DB05676	Apremilast	0.55281	0.60184	0.69188
DB00005	Etanercept	0.54921	0.62822	0.71146
DB06681	Belatacept	0.54872	0.51778	0.60321
DB01281	Abatacept	0.54872	0.51778	0.60321
DB00033	Interferon gamma-1b	0.54712	0.52115	0.54888
DB01017	Minocycline	0.54283	0.6186	0.71123
DB00051	Adalimumab	0.5417	0.60514	0.6815
DB01427	Amrinone	0.53984	0.57615	0.65755
DB01041	Thalidomide	0.53867	0.60462	0.6917
DB09052	Blinatumomab	0.5385	0.52442	0.55796
DB08904	Certolizumab pegol	0.53603	0.59474	0.67143
DB06674	golimumab	0.53603	0.59474	0.67143
DB00065	Infliximab	0.53603	0.59474	0.67143
DB00852	Pseudoephedrine	0.53224	0.61075	0.70126
DB01169	Arsenic trioxide	0.52676	0.59345	0.69408
DB08910	Pomalidomide	0.5146	0.60549	0.68611

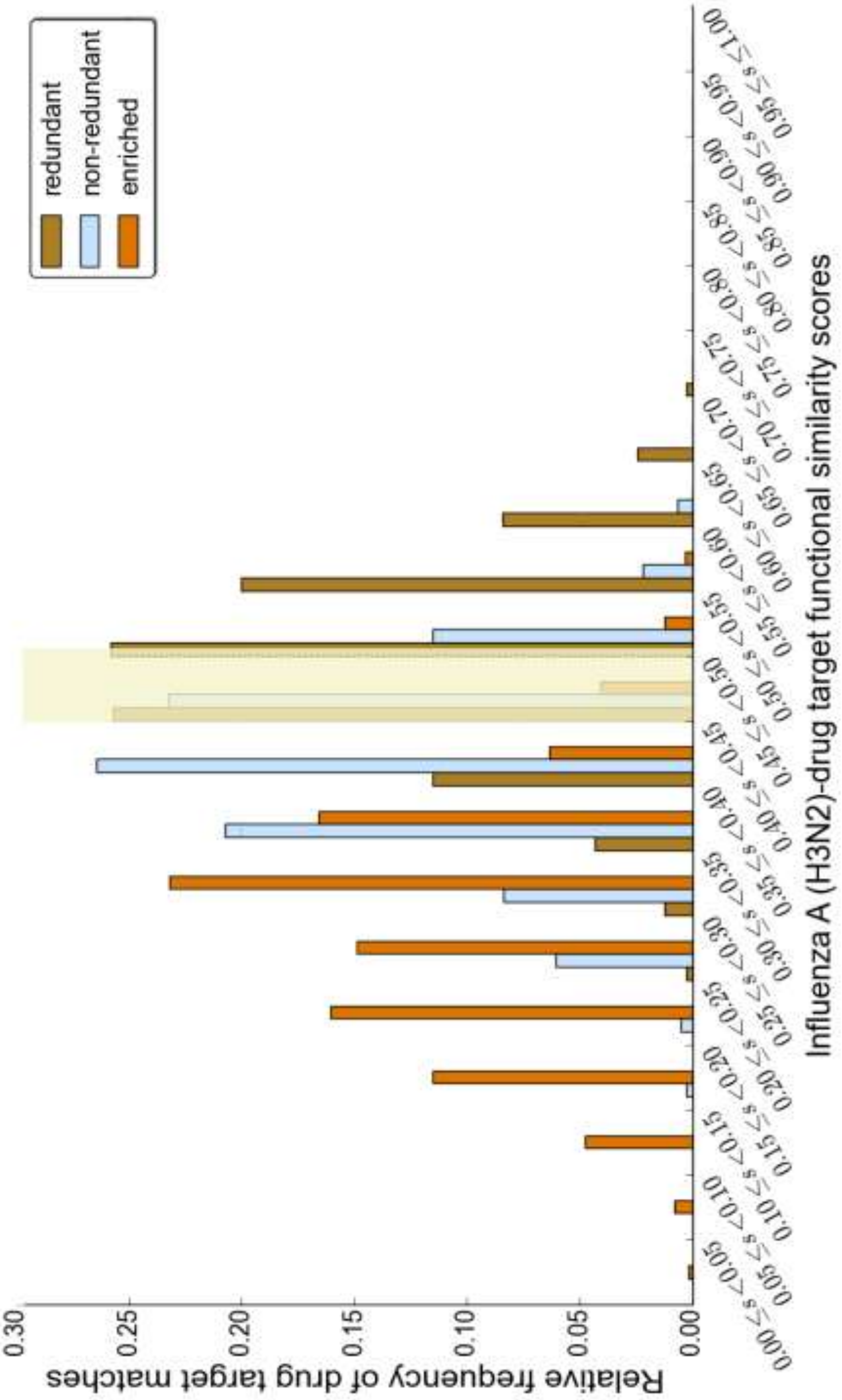

Figure 6: Drug Similarity Scores for H3N2

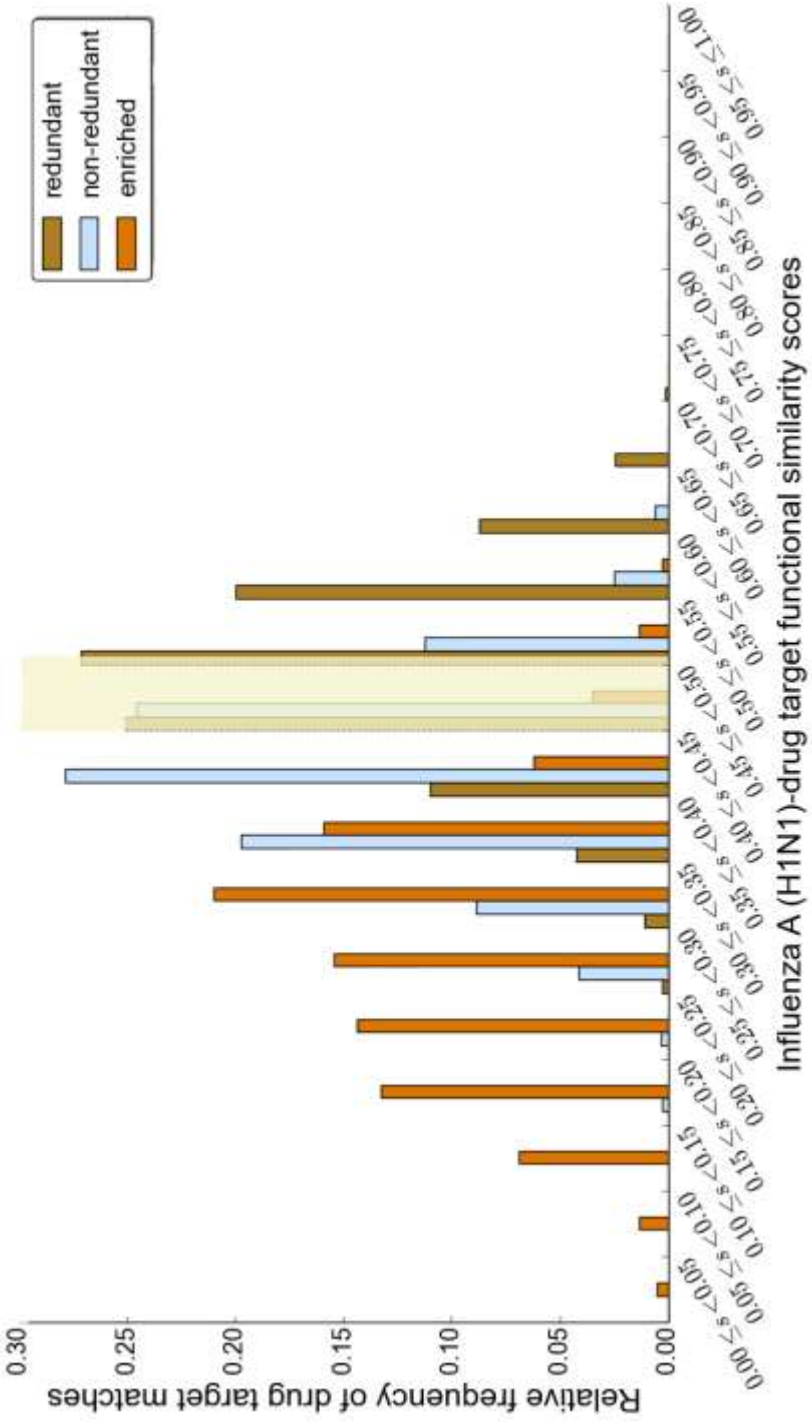

Figure 7: Drug Similarity Scores for H1N1

Summary

In this chapter, the results were discussed. Here, we see results for the many different procedures that were employed in this research work. We see the results for the extraction of differentially abundant proteins where we found two subsets of 69 and 64 out of 542 genes for H1N1 and H3N2, we also identified 9 clusters from which these differentially abundant proteins were found in. We also found the biological processes for which these proteins were in. We then found 273 candidate proteins for H1N1 and 272 candidate proteins for H3N2. These candidate proteins are key proteins that play a vital role and also serve as potential targets for influenza A. Using these candidate proteins, we were able to predict drugs which can be redirected to treat influenza A, total of 71 potential drugs for H1N1 and 83 for H3N2. We also identified the different pathways for which H1N1 an H3N2 belong. The plots and tables provide a detailed description for the results.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview

This chapter presents a total summary on the research work done. Through this work, potential targets of influenza A were found, as well as potential drug targets, which can be very useful to many. The following text gives further details on the research work done.

Summary

The purpose of this thesis was to identify potential targets of influenza A, this is because we identified that influenza A virus has a tendency of re-assorting, this means that there is a need to constantly find new ways of detecting potential targets of influenza A in order to be ready when a re-assorted virus resurfaces. Firstly, we collected data from different sources as discussed in chapter three in order to generate a protein-protein functional network which is seen in chapter four. Secondly, we identified potential targets of influenza A using centrality measures (degree, closeness and betweenness), we then went on to extract differentially abundant proteins using the Pearson's chi squared score and then performed an enrichment analysis and a pathway analysis through KEGG. Finally, we found potential drugs by matching the identified targets using the BMA model to help set a cut off. All of these results are shown in chapter four.

Conclusions

In this work, we have found key proteins from the human network which are potential targets for influenza A. We have gained insight into the molecular interaction between humans and influenza A. We also performed a system level analysis based on biological processes and pathway maps in which the interacting proteins were involved. The biological process analysis of the proteins suggested that human proteins participating in the interactions play a critical role within the human system. The pathway analysis also revealed some proteins are known to play a role in the influenza A pathway and IAV may take over the host to acquire nutrients. We also found drug targets based on the influenza A protein targets, which were used to find new drug targets which can be used to produce new drug indications for influenza A.

In conclusion, we submit that influenza A still needs surveillance, however, the main significance of this work is that it provides a systematic, concise and rapid method of finding medical solution to influenza A and then to other diseases. We believe the methods presented in this work are useful for the future. Simply because, this time round, the analysis is not based on the pathogen but on the host, by trying to identify what exactly (in our case the key proteins) the virus is likely to attack, and used these same key proteins to find new drug indications.

Recommendations

This systems level analysis performed on the influenza A data has provided us with interesting results. Nevertheless, it cannot be concluded that this model will work perfectly for further other diseases. We therefore recommend that this method be extended to other infections to test how effective it will be on those diseases. In future, it will be good to include an analysis from pharmaceutics and drug production to be sure the drugs discovered from the analysis can be used and not just conclude as potential drug that have the possibility of treating a particular disease.

REFERENCES

Agostino, M. (2012). Introduction to the blast suite and blastn. Practical *Bioinformatics*, 47–72.

Anderson, D. R. (2010). Some background on why people in the empirical sciences may want to better understand the information-theoretic methods. pdf Retrieved https://www.jyu.fi/bioenv/en/divisions/ eko/ coeevolution/events/itms/why, 06–23.

Bailey, K. (1994). Numerical taxonomy and cluster analysis, typologies and taxonomies. Sage Atlanta, Ga, 34–65.

Barabasi, A.-L. (2016). Network science. http://barabasi.com/f/625 .pdf. (Accessed:June 2016)

Barab´asi, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Reviews Genetics, 12(1), 56–68.

Barren¨as, F. (2012). Bioinformatic identification of disease associated pathways by network based analysis. Link¨oping Univeristy Medical Dissertations (1326).

Blondel, V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of community hierarchies in large network, 2008. J. Stat. Mech. P, 1008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. Physics reports, 424(4), 175– 308.

Brown, J. S., Hussell, T., Gillilan S. M., Holden, D. W., Paton, J. C., Ehrenstein, M. R., & Botto, M. (2002). The classical pathway is the dominant 118complement pathway required for innate immunity to streptococcus pneumonia infection in mice. Proceedings of the National Academy of Sciences, 99(26), 16969–16974.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., & Apweiler, R. (2004). The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucleic acids research, 32(suppl1), D262–D266.

CDC. (2009). Key facts about swine flu. http://www.cdc.gov/ h1n1flu /pdf/keyfacts.pdf. ([Online; Accessed April 2016])

Chimusa, E., Mazandu, G., & Mulder, N. (2016). Gene ontology semantic similarity tools: Survey on features and challenges for biological knowledge discovery. Briefings in bioinformatics. (In Press)

Chutinimitkul, S., Suwannakarn, K., Chieochansin, T., Damrongwatanapokin, S., Chaisingh, A., Amonsin, A., & Poovorawan, Y. (2007). H5n1 oseltamiv irresistance detection by real-time pcr using two high sensitivity labeled taqman probes. Journal of virological methods, 139(1), 44–49.

Consortium, G. O.. (2010). The gene ontology in 2010: extensions and refinements. Nucleic acids research, 38(suppl 1), D331–D335.

Davis, A. M., Chabolla, B. J., & Newcomb, L. L. (2014). Emerging antiviral resistant strains of influenza A and the potential therapeutic targets within the viral ribonucleoprotein (vrnp) complex. Virology journal, 11(1), 1.

Deng, H.-W., & Shen, H. (2007). Current topics in human genetics: studies in complex diseases. World Scientific.

Fouchier, R. A., Bestebroer, T. M., Herfst, S., Van Der Kemp, L., Rimmelzwaan, G. F., & Osterhaus, A. D. (2000). Detection of influenza A viruses from different species by pcr amplification of conserved sequences in the matrixgene. Journal of clinical microbiology, 38(11), 4096–4101.119

Garten, R. J., Davis, C. T., Russell, C. A., Shu, B., Lindstrom, S., Balish, A., Deyde, V. (2009). Antigenic and genetic characteristics of swine-origin 2009 a (h1n1) influenza viruses circulating in humans. science, 325(5937), 197–201.

Guo, M. (2010). Biological pathways-a pathway to explore diseases (http: //biochem218.stanford.edu/Projects%202010/Guo%202010.pdf.)

Hain, J. (2010). Comparison of common tests for normality. Retrieved from www. statistik-mathematik. uni-wuerzburg. de/.../da hain final. pdf .

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., & Valencia, A. (2004). Intact: an open source molecular interaction database. Nucleic acids research, 32(suppl 1), D452–D455.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. Bmj, 327(7414), 557–560.

Hindiyeh, M., Ram, D., Mandelboim, M., Meningher, T., Hirsh, S., Robinov, J., & Grossman, Z. (2010). Rapid detection of influenza a pandemic (h1n1) 2009 virus neuraminidase resistance mutation h275y by real time reverse transcriptase. Journal of clinical microbiology, 48(5), 1884–1887.

Jiao, L., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., & Lu, Z. (2016). A survey of current trends in computational drug repositioning. Briefings inbioinformatics, 17(1), 2–12.

Leeming, G. H., Kipar, A., Hughes, D. J., Bingle, L., Bennett, E., Moyo, N. A.,. & Sample, J. T. (2015). Gammaherpesvirus infection modulates the temporal and spatial expression of scgb1a1 (ccsp) and bpifa1 (splunc1) in the respiratory tract. Laboratory Investigation, 95(6), 610–624.120

Li, B., Wang, J. Z., Feltus, F. A., Zhou, J., & Luo, F. (2010). Effectively integrating information content and structural relationship to improve the go based similarity measure between proteins. arXiv preprint arXiv:1001.0958.

Lim, B. H., & Mahmood, T. A. (2011). Influenza a h1n1 2009 swineflu and pregnancy. The Journal of Obstetrics and Gynecology of India, 61(4), 386–393.

Liu, Y., Bartlett, J. A., Di, M. E., Bomberger, J. M., Chan, Y. R., Gakhar, L., Di, & Y. P. (2013). Splunc1/bpifa1 contributes to pulmonary host defense against klebsiella pneumoniae respiratory infection. The American journal of pathology, 182(5), 1519–1531.

Mazandu, G., & Mulder, N. (2011a). Generation and analysis of large-scale data-driven mycobacterium tuberculosis functional networks for drug target identification. Advances in bioinformatics, 2011.

Mazandu, G., & Mulder, N. (2011b). Scoring protein relationships in functional interaction networks predicted from sequence data. PLoS One, 6(4), e18607.

Mazandu, G., & Mulder, N. (2013a). Dago-fun: tool for gene ontology-based functional analysis using term information content measures. BMC bioinformatics, 14(1), 284.

Mazandu, G., & Mulder, N. (2013b). Information content-based gene ontology semantic similarity approaches: toward a unified framework theory. BioMed research international, 2013.

Mazandu, G., & Mulder, N. (2014). Information content-based gene ontology functional similarity measures: which one to use for a given biological data type? PloS one, 9(12), e113859.

Meijer, A., Lackenby, A., Hay, A., & Zambon, M. (2007). Influenza antiviral susceptibility monitoring activities in relation to national antiviral stockpiles 121 in Europe during the winter 2006/2007 season. Euro surveillance: bulletin Europeen sur les maladies transmissibles= European communicable disease bulletin, 12(4), E3–4.

Meijer, A., Lackenby, A., Hungnes, O., Lina, B., Van Der Werf, S., Schweiger, B., & Hay, A. (2009). Oseltamivir-resistant influenza virus a (h1n1), europe, 2007–08 season. Emerging Infectious Diseases, Volume 15(4).

Newman, M. E. (2003). The structure and function of complex networks. SIAM review, 45(2), 167–256.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falc˜ao, A. O., & Couto, F. M. (2008). Metrics for go based protein semantic similarity: a systematic evaluation. BMC bioinformatics, 9(5), 1.

Pin˜ero, J., Queralt-Rosinach, N., Bravo, A`. Deu-Pons, J., Bauer-Mehren, A., Baron, M., & Furlong, L. I. (2015). Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015, bav028.

Presanis, A. M., De Angelis, D., Hagy, A., Reed, C., Riley, S., Cooper, B. S., & Lipsitch, M. (2009). The severity of pandemic h1n1 influenza in the united states, from april to july 2009: a bayesian analysis. PLoS Med, 6(12), e1000207.

Rapanoel, H., Mazandu, G., & Mulder, N. (2013). Predicting and analyzing interactions between mycobacterium tuberculosis and its human host. PloS one, 8(7), e67472.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR), 11, 95–130.122.

Reza, F. M. (1961). An introduction to information theory. Courier Corporation.

Rieke, F., Warland, D., & de Ruyter van Steveninck, W., R & Bialek. (1997). Spikes: Exploring the neural code. Cambridge, MA: Massachusetts Institute of Technology, 3.

Schlicker, A., Domingues, F. S., Rahnenf¨uhrer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. BMC bioinformatics, 7(1), 1.

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In Ecai (Vol. 16, p. 1089).

Seddiqui, M. H., & Aono, M. (2010). Metric of intrinsic information content for measuring semantic similarity in an ontology. In Proceedings of the seventh asia-pacific conference on conceptual modelling-volume 110 (pp. 89–96).

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., & Bhatt, S. (2009). Origins and evolutionary genomics of the 2009 swineorigin h1n1 influenza a epidemic. Nature, 459(7250), 1122–1125.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). Biogrid: a general repository for interaction datasets. Nucleic acids research, 34(suppl 1), D535–D539.

Van Kerkhove, M. D., Vandemaele, K. A., Shinde, V., Jaramillo-Gutierrez, G., Koukounari, A., & Donnelly, C. A. (2011). Risk factors for severe outcomes following 2009 influenza a (h1n1) infection: a global pooled analysis. PLoS medicine, 8(7), 964.

Von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). String: a database of predicted functional associations between proteins. Nucleic acids research, 31(1), 258–261.123

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge university press.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., & Hassanali, M. (2008). Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 36(suppl 1), D901–D906.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M. K., Marcotte, E. M., & Eisenberg, D. (2000). Dip: the database of interacting proteins. Nucleic acids research, 28(1), 289–291.

Zhang, P., Zhang, J., Sheng, H., Russo, J. J., Osborne, B., & Buetow, K. (2006). Gene functional similarity search tool (gfsst). BMC bioinformatics, 7(1), 1.

Zhou, Z., Wang, Y., & Gu, J. (2008). A new model of information content for semantic similarity in wordnet. In Future generation communication and networking symposia, 2008. fgcns'08. second international conference on (Vol. 3, pp. 85–89). 124