A systems biology approach- quantification and molecular insights into influenza-A virus infection.

Dissertation

To Fulfill the
Requirements for the Degree of
"doctor rerum naturalium" (Dr. rer. nat.)

Submitted to the Council of the Faculty
of Biology and Pharmacy
of the Friedrich Schiller University Jena

Submitted by: M.Sc. Bioinf., Manchanda, Himanshu

Born on: March 11, 1987 in New Delhi, India

Prof. Dr. Marc Thilo Figge
Applied Systems Biology (ASB),
Hans Knoell Institute, Jena, Germany

PD Dr. Michaela Schmidtke
Institute of Virology and Antiviral Therapie (IVAT),
University Hospital Jena, Jena, Germany

Prof. Dr. Ronald Westra
Biomathematics and Bioinformatics (BMI)
Department of Knowledge Engineering
Maastricht University, Maastricht, Netherlands

Date of Defense: December 18, 2015

”Statisticians, like artists, have the bad habit of falling in love with their models.”
George E.P. Box

Abstract

Influenza virus (IV) circulating worldwide, are highly contagious and can cause acute to severe respiratory disease. Annually, around 500 million individual get infected with influenza, which causes about 500,000 deaths worldwide, including 5000-8000 deaths in Germany. The biological basis for the increased severity of some IVs remains unclear. Recently, it is shown that variability of the amino acid in the 222-position of Hemagglutinin (HA-222) allows evolution of quasispecies in the host organism and is associated with strong inflammation as an important hallmarks of severe infection with pandemic H1N1 influenza A virus of 2009 (A(H1N1)pdm09) (Seidel et al., 2014). modelling influenza kinetics plays an integral role in understanding the differences and pathogenicity of influenza viruses of different virulence. Moreover, investigating the molecular mechanisms of severe A(H1N1)pdm09 is of great importance in controlling the complications and reducing the pulmonary damage.

Together, this emphasises the need to develop a new small-scale mathematical model focussed with only a few state variables describing influenza A virus (IAV) kinetics of different strains. Consequently, categorizing them into mild and severe IAV strains, based on a handful of parameters. For the first time, we did the modelling for dynamics of four different IAV strains and the established model allows quantifying the infection process by using a few interpretable parameters (Manchanda et al 2014). The modelling results also revealed that the late pro-inflammatory response plays some adverse role on the disease condition. Most importantly, the pro-inflammatory response is specific for the virus strain thus therapeutic profile should be design specifically for the virus strains.

After finding the severe A(H1N1)pdm09 strain,A/Jena/5258/2009, we performed a comprehensive analysis of whole genome expression changes of the lung of BALB/c-mice infected with mouse-passaged isolate A/Jena/5258/2009 (mpJena/5258). Surprisingly, most of the genes differentially expressed during mpJena/5258 infection is associates with biphasic gene expression profile. Finally, using a reverse engineering strategy to construct a regulatory network which give us more insight into key interactions leading to severe pathogenesis. The inferred gene regulatory network of influenza infection shows a positive feed-back loop. The gene Ifng coding for interferon gamma is involve in the positive feedback mechanism is associates with the IAV infection, which correlates with evolution of HA-222D/G quasispecies found by Seidel et al 2014., and leads to overwhelming immune response.

Our results show insights into the host factors involved in the severe A(H1N1)pdm09 and show a significant correlation of expression profiles of the genes involved in positive

feedback loop (Ifng, Stat1 and Tlr3) with the clinical symptoms score. However, further experimental investigation with gene expression profiling of the host will need to understand differences in gene expression in mild and severe influenza. These findings play a crucial role in the optimizing the therapeutic profile of drug administration and shall be designed specifically for the virus strain.

Zusammenfassung

Influenzaviren zirkulieren weltweit, sind sehr ansteckend und können schwere akute Erkrankungen der Atemwege verursachen. Jährlich werden weltweit rund 500 Millionen Personen mit Influenzaviren infiziert, mit etwa 500.000 Todesfällen, darunter 5.000 bis 8.000 in Deutschland. Die biologische Grundlage für die Schwere einiger Influenzavirus-Erkrankungen bleibt unklar. Unlängst wurde gezeigt, dass die Variabilität der Aminosäure im Hämagglutinin in Position 222 (HA-222) die Evolution von Quasi-Spezies im Wirtsorganismus ermöglicht und mit starken Entzündungserscheinungen als wichtige Kennzeichen einer schweren Infektion mit pandemischen H1N1 Influenza-A-Viren von 2009 (A(H1N1)pdm09) assoziiert ist (Seidel et al., 2014).

Die Modellierung der Influenza-Kinetik liefert einen wichtigen Beitrag zum Verständnis der Unterschiede und Pathogenitätsmechanismen von Influenzaviren unterschiedlicher Virulenz. Darüber hinaus ist die Untersuchung von molekularen Mechanismen der schweren A(H1N1)pdm09-Infektion sehr wichtig für die Intervention bei Komplikationen und zur Verringerung der Lungenschäden.

Dies unterstreicht die Notwendigkeit, ein neues fokussiertes mathematisches Modell mit nur wenigen Zustandsgrößen zu entwickeln, das den Verlauf der Infektionen mit verschiedenen Influenza-A-Virus-Stämmen im Wirt beschreibt und anhand weniger Parameter kategorisiert in milde und schwere Infektionen. Die Ergebnisse der Modellierung zeigten, dass späte pro-inflammatorische Reaktionen eine schädliche Wirkung auf den Krankheitszustand haben. Wichtig ist, dass die pro-inflammatorische Antwort spezifisch für den Virusstamm ist, weshalb das therapeutische Profil individuell für den jeweiligen Virusstamm gestaltet werden sollte.

Für den in BALB/c-Mäusen hoch-virulenten A(H1N1)pdm09-Stamm (A/Jena/5258/2009) wurde eine detaillierte Analyse durchgeführt, indem die genomweite Änderung der Genexpression in der Lunge der infizierten Mäuse untersucht wurde, die mit dem Maus-passagierten Isolat A/Jena/5258/2009 (mpJena/5258) infiziert wurden. Schließlich wurde mit einem Reverse-Engineering-Ansatz ein regulatorisches Netzwerk konstruiert, das einen Einblick in die wichtigsten genregulatorischen Interaktionen unter den Bedingungen der schweren mpJena/5258-Infektion erlaubt.

Überraschenderweise – und in dieser Arbeit erstmalig – konnte gezeigt werden, dass die meisten differenziell exprimierten Gene Expressionsprofile über den Verlauf der mpJena/5258“-IAV-Infektion haben, die zweigipflig sind. Das genregulatorische Netzwerkmodell zeigt eine positive Rückkopplungsschleife. Im rekonstruierten mathematischen Netzwerkmodell sind die Expressionsprofile der Gene, die in die Rückkopplungsschleife

eingebunden sind, mit der Entwicklung der HA-222D/G-Quasispezies des mpJena/5258 verbunden (Seidel et al., 2014). Es wurde eine signifikante Korrelation der Expressionsprofile der zur Rückkopplungsschleife gehörenden Gene IFNg, STAT1 und Tlr3 mit den klinischen Symptomen gefunden.

Die Ergebnisse der Modellierung geben Einblicke in die wirtsseitigen Faktoren, die für die schwere A(H1N1)pdm09-Infektion charakteristisch und hypothetisch auch verantwortlich sind. Es sind jedoch weitere experimentelle Untersuchung mit Genexpressionsanalysen des Wirtes erforderlich, um Unterschiede in der Genexpression bei einer milden und und schweren Influenza zu verstehen. Diese Erkenntnisse könnten eine wichtige Rolle für die Optimierung der antiviralen Therapie mit einem IAV-Stamm-spezifischen Arzneimittel spielen.

Acknowledgements

I would like to dedicate this thesis to my mother, Vijay Laxmi, with love and gratitude for her encouragement to pursue the highest level of studies and research.

Although my name appear on first but there have been substantial non-textual contribution from a lot of people who made this thesis possible. First of all, I would like to convey my heartfelt thanks to Prof. Dr. Reinhard Guthke and PD Dr. Michaela Schmidtke who have together supervised me throughout the study and initiated this highly interesting project. I specially would like to mention their continuous encouragement and their belief in my abilities to achieve high standard goals. I believe that I have grown professionally and personally with the mentorship provided by both of them.

I wish to express my deep sense of gratitude to PD Dr. Michaela Schmidtke's group (Anja, Martina, Lilia, Elisabeth, Zhongli) Institute of Virology and Antiviral Therapy, University Hospital Jena, especially Nora Siedel for having fruitful discussions, sharing knowledge on influenza virus infection and in research collaboration.

I am especially grateful to my colleagues who with their feedback, inputs and coffee-chats made working enjoyable. I cherish the discussions and the time spent with Fabian, Sebastian Vlaic, Joerg, Katya, Robert, Steffen, Sylvie, Andreas, Uwe, Patricia, Thomas, Mathiew, Sebastian Henkel, who in time erased the boundaries between colleagues and friends and shifted to later.

I owe a special thanks to Fabian, Nina, Judith who have taught me various sides of German culture.

I am deeply indebted to friend group named Santo's (Naren, Jitu, Susu, Divlu, Sonika, Manmee, Varu, and Neetika) and Three musketeers (Iron kirti and Sathish) in Jena and would like to pay a big hug for all of them together for fun and light moments. A special thanks to my childhood friend, Late Gaurav Gulleria (1986-2015), who will always stay in my heart.

I also would like to thank my best buddies Sudipto, Sanjay, Akhil, Manish, and Vicky who have always trusted that whatever I was doing was something special and is going to help scientific community sooner or later.

I also would like to thank Evi, Laura, Eike and Kay who have been my teammates and coaches in USV, Jena Badminton team.

Last but not the least, I would like to thank my best friend cum wife Neetika Nath for her support, encouragement, quiet patience and unwavering love during the past few

years. I thank my Mother and my younger Brother, Geetanshu Manchanda, for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on. Also, I thank Neetika's parents, they provided me with unending encouragement, support and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Influenza Virus

Influenza viruses (IV) circulating worldwide are highly contagious and can cause acute to severe respiratory disease. Influenza viruses are remarkable in nature because of their high mutability [1, 2] and high frequency of genetic reassortment and resulting antigenic variations in the surface glycoproteins, Neuraminidase and Hemagglutinin. Annually around 500 million individuals are infected by influenza virus, which cause about 500,000 deaths [3, 4] worldwide, including 5000-8000 deaths in Germany [5]. Influenza type A is antigenically highly variable and is mainly responsible for most cases of epidemics and pandemics.

1.1.1 Types and Nomenclature

The influenza viruses which belong to the family of Orthomyxoviridae, are characterized by segmented, negative-strand RNA genomes. There are three types of Influenza viruses, namely Influenza A, B and C viruses, of which only type A, and B viruses are most significant as pathogens whereas type C virus is less common and only causes mild disease in children [6]. Sequencing has confirmed that these viruses share a common genetic ancestry [7] and electron microscopy reveals that influenza A, and B virus are virtually indistinguishable.

Influenza A virus (IAV) is further characterized by their subtype of their surface glycoprotein, the hemagglutinin (HA) and the neuraminidase (NA). Many genetically distinct subtypes, 18 for HA (H1-H18) and 11 for NA (N1-N11) [8], have been found in the circulating influenza A virus. Many subtypes are specific to birds, some to swine but not

all the resulting IAV subtypes infect humans. Only three HA (H1, H2, H3) and two NA (N1 and N2) subtype's combinations are commonly found in humans [8].

Influenza viruses have a standard nomenclature that includes virus type, species from which it was isolated if on non-humans, location at which it was isolated, isolate number, isolate year and for IAV virus only, HA and NA subtype. For example, A/Jena/5258/2009 (H1N1) was isolate number 5258 of human Influenza A taken in the city of Jena in 2009 has an HA subtype 1 and NA subtype 1 [9].

1.1.2 Structure and Composition

Influenza virus particles are usually spherical and about 100 nm in diameter, although the filamentous form also exists and the length can go up to 300 nm. Both influenza A and B virus genome consisted of eight negative-sense, single stranded viral RNA (vRNA) segments while influenza C has seven segment genome – lacking the neuraminidase gene. Sizes and polypeptide assignments are known for the segments (Table 1.1). Most of the segments code for a single polypeptide protein. The complete nucleotide sequence is known for many influenza viruses. Interestingly, first 12-13 nucleotides at the end of each genomic segment are conserved among it's all eight RNA segments and are very important in viral transcription [6, 10].

TABLE 1.1: The genomic segments of influenza A virus and their encoded proteins.

Segment	Segment length (# of nucleotides)	Encoded polypeptide(s)	Length of polypeptide (in amino acids)	Function
1	2341	PB2 (Polymerase basic 2)	759	Polymerase subunit; mRNA cap elongation
2	2341	PB1 (Polymerase basic 1)	757	Polymerase subunit; RNA elongation; endonuclease activity
		PB1-N40	718	
		PB1-F2	87	Pro-apoptotic activity
3	2233	PA (Polymerase acidic)	716	Polymerase subunit; protease activity
4	1778	HA (Hemagglutinin)	550	Surface glycoprotein; mediated virus attachment to cells; activated by cleavage; fusion activity at acidic pH
5	1565	NP (Nucleoprotein)	498	Associated with RNA and polymerase proteins; helical structure; nucleocapsid
6	1413	NA (Neuraminidase)	498	Associated with RNA and polymerase proteins; helical structure; nucleocapsid
7	1027	M1 (Matrix protein 1)	252	Major component of virion; lies inside of envelope; involved in assemble; interacts with RNPs and NS2
		M2 (Matrix protein 1)	97	Integral membrane protein; ion channel; essential for viral uncoating
8	890	NS1 (nonstructural protein 1)	230	Interferon antagonist protein; inhibits pre-mRNA splicing
		NS2 (nonstructural protein 2)	121	Minor component of virion; nuclear export of viral ribonucleoproteins (RNPs)

Influenza A virus encodes for twelve polypeptides, out of which only two of them are non-essential outside the laboratory settings [11, 12]. The nucleoprotein (NP) associates

with the viral RNA (vRNA) to form a ribonucleoprotein (RNP; 9 nm in diameter), that assumes a helical configuration and forms the viral nucleocapsid. Three large proteins, e.g. PB1, PB2, PA, are bound to the viral RNP and are responsible for RNA replication and transcription. The matrix (M1) protein, which forms a shell underneath the viral lipid envelope is important for its morphogenesis and is the major component of virion, about 40% of viral protein.

The virus particles are surrounded by the lipid envelope which is derived from the cell. The Two glycoproteins, encoded by the vRNA, HA and NA are inserted into the envelope and are exposed as spikes on the surface of the particle. These two surface glycoproteins are important antigens, which determine the antigenic variation of influenza viruses and host immunity. The HA represents about 25% of the viral protein which helps the virus to recognise N-acetylneuraminic (sialic) acid on the host cell surface [10]. The NA represents about only 5% of the total protein but are very important for viral release because of their sialidase activity. The M2 ion-channel protein as well as NS2 (nuclear export protein; NEP) are also present in the envelope but at only few copies per particle [6, 10].

1.1.3 Transmission

The transmission of the disease-causing influenza virus occurs in three ways: 1) by direct contact with infected individuals; 2) by contact with the contaminated objects; 3) by airborne infection, aerosols [8, 13]. The influenza virus shedding begins the day before symptoms appear and efficient virus shedding to continues until five to seven days. Nasal secretions, which contain virus particles, are responsible for transmission by direct contact or by contaminated objects. An infected person will frequently touch their nose or conjunctiva, placing the virus on the hand. Intimate or non-intimate contact (e.g. shaking hands) will transfer the virus to another person, who will then infect themselves by touching their nose or eyes [8].

Respiratory transmission depends upon the production of fine aerosols that contain virus particles. Coughing from an infected individual may produce several hundred droplets, a good sneeze can generate up to 20,000 particles [13, 14]. Droplets of 1-4 microns in diameter are called *droplet nuclei*, these stays in the air for very long period of time and these droplets nuclei not only travel long distances, but can also reach the lower respiratory tract. Inhalation of droplets and droplet nuclei places a virus in the upper respiratory tract, where they initiate infection. Influenza transmission can be reduced by covering nose and mouth when coughing or sneezing, and by washing hands often with soap and water or alcohol-based hand cleaners. [8, 13, 14]

1.1.4 Symptoms

Symptoms of classical influenza usually appear between 18 and 72 hours post infection and include chills, headache, and dry cough, followed by fever. Other symptoms includes muscle and joint aches, sore throat, runny nose, nausea and anorexia [8, 10].

The virus can spread throughout the entire respiratory tract and thus, the bronchi, the lower trachea and the lungs become infected. The degeneration and deaths of infected cell are a prerequisite for the bacterial super-infection. This usually occurs after two to three days after the onset of influenza infection and can significantly prolong the course of respiratory infection. The most common bacteria that cause super-infection are *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* [15, 16]. Influenza infections can also be asymptomatic or only mild symptoms like the running flu while the other extreme are very severe cases, to death.

1.1.5 Seasonality, Epidemics and Pandemics

Influenza viruses are remarkable in nature because of their frequent antigenic changes that occur in HA and NA. The two surface antigens or surface glycoproteins of influenza undergo antigenic variations independent of each other. Minor antigenic variations are called as **antigenic drift** (characteristic of all three influenza viruses) whereas major antigenic changes in HA and NA are called as an **antigenic shift** (characteristic feature of only influenza A virus) which results in the appearance of a novel subtype. Antigenic shift is most likely ended up in an epidemics or even pandemics. Antigenic drift is due to the accumulation of point mutation in the gene, which results in changes in amino acid of the protein. Antigenic drift reflects drastic changes of the viral surface glycoproteins due to reassortment of viral genome segment. The segmented genomes of influenza virus reassort readily in double infected cells, for example cells infected with H1N1 and H3N2.

Typically, influenza epidemics occur abruptly, reaching their peak within two to three weeks and last for approximately five to six weeks. Prolonged epidemics can occur by either successive or overlapping waves of infection of different virus types. The pandemics, i.e. global epidemics with high morbidity and mortality, may occur when a new virus subtype or variant is formed by the antigenic shift against the population lacking immunity against it. In 20th century, globally there were three influenza pandemics caused by genetic modification of influenza A virus, this happened in 1918-19 (H1N1) [17], 1957-58 (H2N2) [18], 1968-69 (H3N2) [19]. The most severe pandemic was in the winter of 1918-1919, the so-called Spanish influenza (A/H1N1), killed around 40-50 million people [20]. The biological basis for the increased severity of some influenza viruses remains

unclear. Unpredicted mutations which leads to intra-host evolution of quasi-species [2], and strong inflammation are important hallmarks of severe pandemic influenza infection [1, 11, 21–23].

1.2 Mathematical Modelling

In science, a model is generally considered either as a part of a or a representation, of a system that we aim to study, while modelling is an attempt to describe a better understanding of the elements of the systems, their states and their interaction with other elements. A mathematical model represents the examined systems (dynamics, for instance) in the form of various types of equations, often consists of differential equations, which intend to understand the system. The model should be defined precisely to include key elements, which can, in principle, be used to simulate the behaviour of the system on a computer. For example, a model can be used to describe relatively low-level molecular biology processes like transcription, translation, gene regulation, cellular signalling and so on. Moreover, it can be used at higher level describing the behaviour and evolution of population of individual organism. In addition, mathematical modelling plays a far more important role, especially in the studies of dynamics of infectious agents and have proven to be useful tools in the analysis of viral infections [24–31]. For instance, the dynamics of human deficiency virus (HIV) infections were poorly understood until simple mathematical models were developed. These models could help in estimating the rate of HIV replication, the number of virus particles produced and cleared daily, and the average life span of productively infected $CD4^+T$ cells [25, 26].

Emergence of new influenza virus strain via genome reassortment or the emergence of antibiotic-resistant strain of influenza, both keep a general interest in infectious diseases and their control going. Mathematical modelling of infectious disease has a long history in mathematical biology. The first mathematical model that could be used to describe an influenza virus dynamics within a single infected host dated back in 1976 [24]. A compartmental model was used to describe the dynamics of influenza virus within infected mice, further details about mathematical modelling and their advancement can be found in the next sections.

1.2.1 State Variables and Model parameters

The primary components of a mathematical model correspond to the elements involved in the systems. The abundance of each element is assigned to a **state variable** within the model. The collection of all these state variables is called as the state of the system. It

gives us a complete description of the events or system's condition at any particular time. The model's dynamics behaviour is the time course for the collection of state variable. Besides variables of state, mathematical models also include **parameters**, whose values are sometimes fixed but can be estimated with various optimization algorithms [32]. Model parameters define the interaction between the elements of the systems and with the environment. For example, model parameters are: degradation rate, replication rate, carrying capacity of a pathogen and so on. A change in the value of a model parameter corresponds to a change in the environmental or in the systems itself. As a result, model's parameters are typically held constant during model simulation. These values can be varied to explore systems behaviour under different experimental conditions.

1.2.2 Linearity and Non-linearity

A relationship is called as **linear** if it is of a direct proportionality or relationship. For example, the variable x and y are linearly related by the equation $x = ky$, where k is a fixed constant. Linearity allows for an effortless extrapolation, which means doubling of x will lead to doubling of y regardless of their values. A dynamic mathematical model is called as linear if all interactions among its components are of linear relationships. However, sometimes it is a highly restrictive condition, and as a result they display only a limited range of behaviour. For example, the rate of replication or growth of bacteria in a cell culture medium is directly proportional to the continuous availability of resources in the medium.

Any relationship that is not linear is referred to as **non-linear**. Non-linear relations need not to follow any specific pattern and so are difficult to address any generality. The non-linearities that appear most often are saturations, in which one variables increases with another variable at a diminishing rate, so that the dependent variable tends to a limiting value. Two kinds of saturating relationship that occur most often are *hyberbolic saturation*, in which the rate of the increase of one variable declines continuously as the value of other variable increases. Secondly, the *sigmoidal saturation*, where one variable initially grows slowly with another variable, then passes through a phase of rapid growth before saturating.

1.2.3 Stochastic and Deterministic Models

There are two types of models that are useful in the study of the infectious diseases at the population scale: these are stochastic and deterministic models.

Stochastic models rely on chance variation in risks of exposure, disease, and other factors [33]. They give us much more insight into an individual-level modelling, such as, taking into consideration small population size where every individual play a significant role in the model. Therefore, they are used when known heterogeneities are important as in small or isolated populations. Stochastic models have several advantages. More specifically, they can allow close watching of each individual in the population on a chance basis. However, sometimes they can be more laborious to set up and need many simulations to yield useful predictions. These models can become mathematically very complex, take more computational time and do not contribute to an explanation of the dynamics.

However, *Deterministic models*, which are also called as a compartmental model, attempts to describe and explains what happens to the mean at the population scale instead on an individual scale. They fit well to large populations. These models categorises individuals into different sub groups/compartments, such as, SIR model consists of three compartments represented by Susceptible, Infectious and Recovered [34]. The infection or the attack rate is defined as the best transition rate, how the size of one compartment changes with respect to the other, which measures the rate at which susceptible becomes infected.

Most of the mathematical model which have been used, until now, to describe the behaviour of infectious disease are deterministic as they require less data, relatively easy to set up, many computer software's are readily available and user-friendly. Therefore, deterministic models are commonly used to explore whether a particular control strategy will be effective or not, because of the well understanding of the SIR model. Furthermore, many other more complex models exist that can incorporate stochastic elements. However, in my thesis, I will focus more on the deterministic models.

1.2.4 Basic Viral Dynamics Model

Perelson [27] has described a basic model, which has been used to study human immunodeficiency virus (HIV), hepatitis C virus (HCV), hepatitis B virus (HBV). The basic model of viral infection dynamics was based on the classic SIR model. Target cells (T), which are susceptible to infection are infected by the virus (V) to become infected I by a constant rate of κ. Whereas, target cells are assumed to be continuously produced from a source by a constant rate of λ and these cells have a natural death rate at d per cell.

Each infected cell is assumed to produce new virus particles at a constant average rate ρ per cell. The newly produced virus particles can either infect new healthy cells at a rate

κ as explained above or get cleared from the body at rate c per virion, which corresponds to a half-life $t_{1/2} = ln(2)/c$. An infected cell has an average lifespan of $1/\delta$, where δ corresponds to the death rate of infected cells, thus, producing an average of $N = \rho/\delta$ virions during its lifetime. Fig. 1.1 shows a schematic diagram of these dynamics, and the associated ordinary differential equations are given Eq. 1.1.

$$\begin{aligned} \frac{dT}{dt} &= \lambda - dT - \kappa VT && (1) \\ \frac{dI}{dt} &= \kappa VT - \delta I && (2) \\ \frac{dV}{dt} &= \rho - cV && (3) \end{aligned} \tag{1.1}$$

However, the classic model of influenza dynamics has been modified to a more accurate representation of the infection dynamics as influenza infections are quite rapid, lasting approximately 7-10 days. The eclipse phase, i.e., the time it takes for viral production to begin after intially becoming infected, was incorporated into the models by including a second class of infected cells (I_1) that are in eclipse phase and cannot yet produce virus. One class of these models are based on the assumption that the cells in the eclipse state (I_1) transformed to a productive state (I_2) at rate k and no cell death occurs before the initiation of virus production. While in other classes, the time spent in the eclipse phase is described by either a fixed or distributed delay [29, 35]. In general, including the eclipse phase does not change the model behaviour and may not be statistically justifiable since the number of parameters are increased [36]. However, when taking these dynamics into account, the parameter estimates made by fitting the model to data enter more biologically realistic ranges [28, 29] and growth kinetics are more accurate [30].

The virus lost from entry into cells is typically ignored in viral kinetic's models, as shown above. Since, this is usually a negligible quantity in comparison with the amount lost by other processes such as phagocytosis. Furthermore, in influenza infection, particularly, it is typical to measure only infectious virus. Therefore, loss of viral infectivity can be modeled as clearance of infectious virus. This has been shown by Beauchemin et al [29], where the parameter estimated using viral titres from infection of MDCK cells in vitro suggested that the loss of viral infectivity is much smaller than the loss of virus from cell entry.

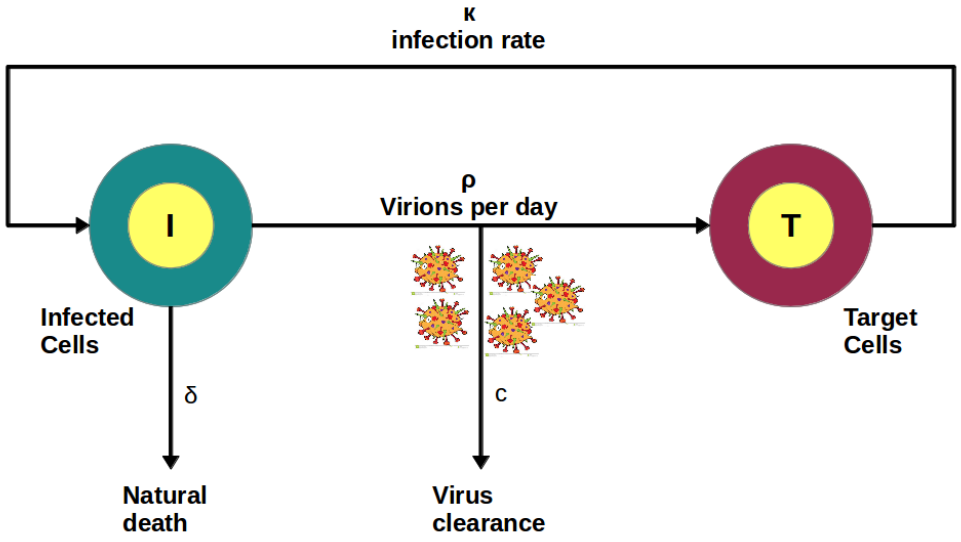


FIGURE 1.1: A basic viral infection dynamic model proposed by Perelson [27]. The model variables T (Target cells), I (Infected cells) and V (Virions) as well as their relations between them are shown in Eq. 1.1.

1.2.5 Application of Mathematical modelling for Understanding Influenza Kinetics

Characterising influenza infection kinetics using data fitting and models to estimate parameter is an effective technique, but obtaining an accurate model in which all parameters are identifiable is challenging. The classic model as discussed above in Fig. 1.1 has 3 equations and 8 parameters for which nonlinear least squares or the maximum likelihood analog are the most frequently used methods to establish parameter values. However, not all the parameters are identifiable [32] and it has been necessary to fix some parameters or restrict the ratios of parameters to ensure they take on biologically relevant values. Modelling and data fitting techniques allow estimation of parameters characterising viral infection. At the same time, model structure can be wrong, error in parameter estimation can occur and different fitting methods can yield distinctive parameters estimates [37, 38], and also multiple solutions because of local optima might confound non linear least square or maximum likelihood fitting.

A number of mathematical models have been proposed, which are dealing with the dynamics of influenza viruses and have been proven to be useful tools in the analysis of influenza and other viral infections [22, 24–28, 31, 39–42]. In contrast, the extremely fast and relatively short duration of replication of influenza A virus in immunocompetent adults invites the search for alternative views of influenza virus and immune system dynamics [28]. Many mathematical models have been developed for influenza virus, where viral titres (viral factors) [24, 28] and/or measurements of at least, one of the

immunological components of hosts [22, 31, 39–42] , like interferons, macrophages, NK cells, B and T cells, were used as experimental data to model the influenza kinetics in different hosts. A study by Larson [24] used a compartmental modelling approach to describe the influenza virus dynamics. However, they did not consider any immune population or factors, so changes in the viral population cannot be connected with any immune response.

The first mathematical model of influenza kinetics to take experimental data into account was proposed by Baccam et al. [28]. Data was taken from experimental infection of volunteers with a seasonal strain of influenza virus, which follows the typical course of an uncomplicated influenza infection, A/Hong King/123/77 (H1N1). Nasal washes were collected daily and serially cultures in 10-fold dilutions to determine viral titres. The simple mathematical model used to describe the kinetic of infection is given by Eq. 1.2:

$$\begin{aligned}
\frac{dT}{dt} &= -\beta TV \\
\frac{dE}{dt} &= \beta TV - kE \\
\frac{dI}{dt} &= kE - \delta I \\
\frac{dV}{dt} &= \rho I - cV
\end{aligned} \tag{1.2}$$

Where susceptible cells, T, can be infected by virus, V, becoming infected cells, I at a rate β. Infected cells produce virus at a rate ρ and die at a rate δ, while virions are cleared at a rate c. Infection is initiated by introducing an initial inoculation of virus V_0 to a population of susceptible target cells T_0. The main limitation of the model was that the effect of immune systems was not explicitly accounted for the virus clearance. Therefore, the model was target-cell limited, which means once the infection has spread. It will only stop until population of target cells has been entirely eliminated by the infection.

Recent models have used approached similar to Baccum et al [28], in order to capture the dynamics of influenza infection using experimental data. Models have been developed to include the effect of the immune response to an influenza infection by comparing models to experimental viral load data and immune response data [22, 31, 39–42].

Another mathematical model attempted to fully characterize the behaviour of within-host influenza infection by taking into account, a substantial immune response [31].

The model consists of 13 variables, representing the virus, epithelial cells and various immune responses, and the behaviour is characterized by more than 60 parameters. Parameter values were estimated from experimental data, and an attempt was made to account for experimentally known infection dynamics, such as estimates of immune system activation, proliferation and decay rates at various times post-infection. However, the results of the model were not compared with experimental viral load or immune response data. Isolating the role of specific immune effects also proved difficult, due to the large number of parameters. Hancioglu et al., [39] developed a model consisting of 10 equations and 27 parameters, but the results were also not directly compared to experimental data due to the difficulty and cost in procuring such extensive viral and host immune response data. Moreover, the model was too complexed and needs more data to estimate the parameters. Pawelek [42] fitted the within−host viral dynamics model to the data derived from six ponies infected with an influenza virus A/eq/Kildare/89 (H3N8). They showed that both innate and adaptive immune response is needed for the viral clearance.

In a review by Smith and Perelson [36], they mentioned about the increasing number of influenza kinetics models based on viral load data and also showed concern about using such data as the only indicator of disease severity. However, immunopathology also plays some role in severe infection [1, 21, 43–46]. For example, Kumar et al., [46] has shown that a strong late immune response can increase the risk of persistent inflammation even after clearing the pathogen. Interestingly, Smith and Perelson [47] suggest that assigning a symptom score throughout the infection could offer a new perspective into the characteristics of an infection. A symptom score is easily attainable and reflects how sick a host is [48].

Both small−scale mathematical models [28, 49] and complex models with larger than 10 equations with larger than 50 parameters [24, 39, 50] have been proposed for studying influenza A infection dynamics. However, complex models require an increased number and quality of experimental data to calibrate the model parameters. Small−scale models have the advantage to be applicable for quantitative comparison of a (also large) set of experiments (e.g., with different virus strains and/or therapeutic strategies) with a limited experimental effort.

Furthermore, studying influenza dynamics *invitro* have strengthened the understanding of viral production and clearance rates. Understanding the differences in the pathogenicity of virus strains is important aspects of influenza biology. Modelling influenza kinetics played an integral role in teasing apart the differences and potential mechanism of a virulent influenza strain compared to a less pathogenic one [47]. However, until now, none of the models have been used to fit the data to quantify different influenza viruses. In

my thesis, I have developed a small-scale mathematical model by using the symptom score as an easily attainable data source. The small-scale model could fit the data of four different influenza virus strains and able to quantify them based on their pathogenicity of infection from severe to mild to weaker. Virus pathogenicity, antiviral immune defence, and pro−inflammatory response were considered as the main model variables in a three-dimensional differential equation model.

1.3 Gene Expression Microarrays

1.3.1 Overview of DNA and RNA

Deoxyribonucleic acid (DNA) encodes the information required for the development and function of a living organism. The structure of DNA is remarkably simple, being formed of a long chain of smaller units (nucleotides) joined together. Each nucleotide can have one of four bases attached; Adenine (A), Cytosine (C) Thymine (T) or Guanine (G), however, in case of Ribonucleic Acid (RNA) has a Uracil (U) base instead of T. The order in which these bases occur in a DNA and RNA molecule is known as the DNA and RNA sequences respectively.

Structurally, a DNA molecule takes the form of a double helix arranged by two strands of DNA. This structure is held together by strong hydrogen bonds between the two strands. The bonding takes place in such a way that A base-pairs with a T base whereas C base-pairs with G, in the opposite strand. This is known as the base complementarity property of DNA and effectively means that the sequence of one strand can be determined by the other, a fact that is exploited during DNA replication.

The entire DNA sequence of an organism is known as its genome. The human genome is estimated to have 3.3 billion bases and can be found in the nucleus of every cell in the body. Rather than being one long DNA molecule, a genome is divided into continuous regions of DNA known as chromosomes. There are 24 chromosomes in the human genome, which are numbered from 1 to 22 plus the sex chromosomes, X and Y. Therefore, the normal healthy cells in the human body contain 46 chromosomes, i.e., two copies of 1 to 22 chromosome (each copy from one parent) and either X and Y for males or X and X for females. Each chromosome is divided into stretches of DNA called genes, which encode the instructions to produce a particular protein. The estimate number of protein-coding genes that the human genome has around 20,000 [51]. However, rather than being one continuous sequence of genes, there are many gaps in the chromosome that are not genes and hence do not code for proteins. In fact, only an estimated 5 to

10% of our genome is used to code for proteins. The purpose of remaining "junk DNA" is a source of much debate, but increasingly is considered to have regulatory function.

The information encoded in the DNA sequence are stored in the nucleus and must be transported to the cytoplasm, where specialised molecules called ribosomes help produce the required proteins. Therefore, sections of DNA are copied i.e. transcribed into messenger RNA (mRNA) molecules that contain the same information as DNA (as in complementary DNA), but in a slightly different form. The main differences are that mRNA is single-stranded, degrades quickly. The entire sequence of each gene is transcribed, even though not all parts of the sequence take part in coding for proteins. Such non-coding regions, known as introns, are removed by splicing before the process of translation starts. Translation uses mRNA as a template to assemble previously synthesised amino acids in the correct order to make particular proteins, with groups of 3 successive bases used to specify the amino acid located at that position in the chain.

Although each cell contains copies of the same genome, the cell will require different combinations as well as distinctive quantity of protein in order to perform its activity within the body. Therefore, the genes that control the production of these proteins may be turned on or off to varying degrees, depending on the internal and external stimuli. These changes confer unique properties to each cell type. The expression level of a gene refers to the amount of mRNA that is made from the DNA template and finally translated into the required proteins.

1.3.2 Gene Expression Microarrays

A microarray, sometimes called as an array, is an experimental technique for simultaneously measuring the expression level of genes ranging from few hundred genes to whole genome. The technology makes use of the base-complementarity property of DNA and the fact that single-stranded mRNA is produced in order for a particular gene to be expressed. Thus, by measuring the amount of mRNA produced we can infer the expression level of the genes involved.

Microarrays are typically constructed by attaching single-stranded DNA sequences, known as probes, to a surface such as a glass slide. Each probe is complementary to the DNA sequence of a particular gene of interest and is placed in spots (or features) at pre-defined locations. Single stranded mRNA from a sample of interest (called the target) is isolated, converted into single stranded DNA (cDNA) and then transcribed into cRNA. These cRNA are next fluorescently labelled, and exposed to the microarray surface. The target RNA then hybridises to its complementary probe sequence on the microarray, whereas non-complementary sequences should fail to hybridise.

The amount of fluorescence observed at each feature can therefore be used to determine the level of expression for each of the genes represented on the array. In the earliest microarrays [52], each feature on the array corresponded to a different gene of interest. However, subsequent developments in microarray production have allowed the same gene to be represented multiple times, thus providing more reliable expression estimates.

Two colour microarrays are used to compare two samples, i.e. control versus treated, on the same microarray. The RNA from the two samples is extracted separately and fluorescently labelled with different dyes, usually red (Cy3) for treated and green (Cy5) for control. Therefore, after hybridisation, each feature is a mixture of red and green fluorescences. A completely red or green feature represents that the particular gene is expressed in one sample, but not in other. Practically, the mixture of red and green fluorescence features are not so clear cut that the statistical methods are required to quantify the contributions of each colour. A *differential expression* analysis aims to figure out, which genes have significantly different expression levels between different experimental conditions under investigation. Such genes are said to be differentially expressed genes (DEGs). See Fig. 1.2 for schematic representation of a typical two colour microarray experiment.

On the other hand, **single-channel** microarray can also be produced to measure the absolute expression level of every gene of interest in a particular sample. Therefore, the fluorescence intensity of each feature is a measure of the expression level of a particular gene. The most popular single-channel microarray technology was that of Affymetrix [53]. These arrays use 25 base-pair probes that are synthesised onto the array surface and each gene is interrogated by a collection of probe pairs, usually 11-20, and are known as probe set.

Microarrays have become an invaluable tool for medical research [54] and provide a great wealth of data that was previously unobtainable. Additionally, microarrays are manufactured for application other than genome expression. For instance, microarrays can be used to interrogate genome-wide association studies like Single Nucleotide Polymorphism (SNP) and structural variation like methylation, which alter the structure of DNA but not the sequence.

1.3.3 Illumina Bead-based Microarrays

In my thesis, I will concentrate on the BeadArray microarray technology which is developed by Illumina, and is becoming widely used and offers many potential benefits over other technologies. Rather than attaching probes onto a microarray at known locations,

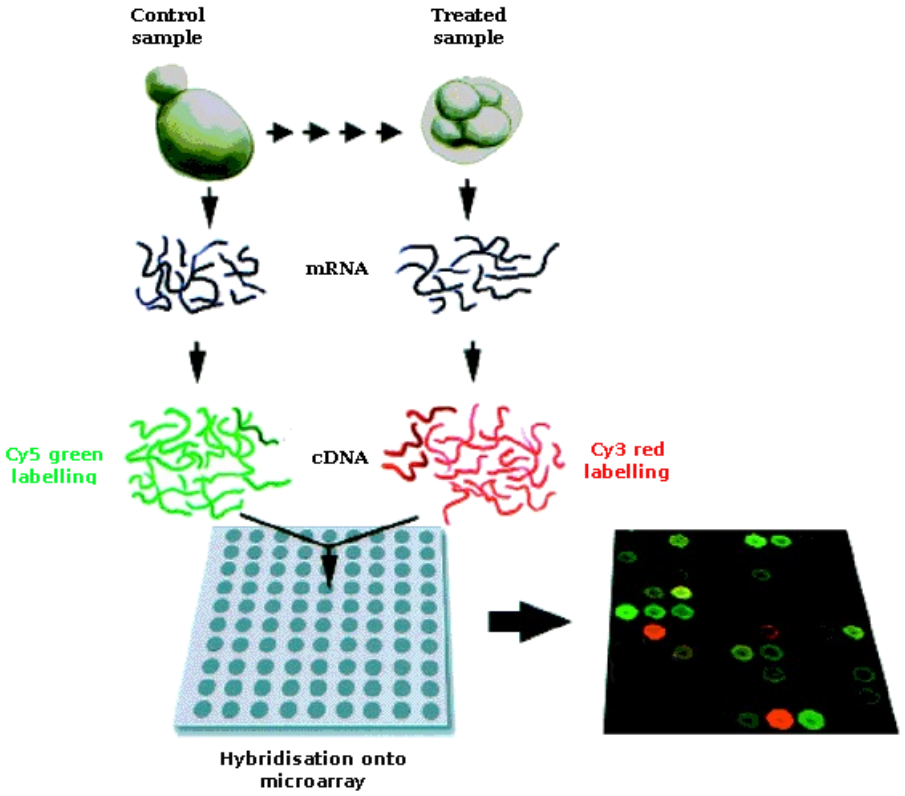


FIGURE 1.2: A typical two colour microarray experiment to compare mRNA expression between a diseased condition to that of a normal condition.

BeadArrays are self-assembling arrays of minute beads with probes attached. Each array is produced separately by exposing an array surface, either a planar silica slide or fibre-optic bundle, to a large collection of pre-prepared beads. This causes the beads to be randomly sampled and assembled into the microwells on the surface of the array slide [55]. When randomly assembled at the slide surface, the beads have a uniform spacing of ≈ 5.7 microns. Each bead is 3 microns in diameter and covered with a specific type of oligonucleotide in many hundreds of thousands copies. Both the number and location of the replicates for the same bead type is random on an array [56]. Therefore, an extra label sequence, an Illumina code, is attached to each bead for decoding [57], with beads of same type also having the same Illumina code. Each Illumina code is designed to hybridise in a predictable way to a series of specially designed dye-labelled sequences. After each hybridisation, each bead is assigned to either of the two states, i.e., red/treated or green/control, depending on the amount of hybridisation, a binary

sequence is determined for each bead. This binary sequence should then uniquely correspond to the predicted response of an Illumina code. These decoding hybridisation are done by Illumina, in order to get that no array will be supplied to the user with a bead type that has less than five replicates.
Along with the high degree of replication within an array, Illumina also offers the capability of processing BeadArrays in parallel, making this technology highly desirable for high-throughput experiments. A Sentrix BeadChip is a glass slide that allows a very high number of observations to be made for a particular sample. Depending on the chip configuration, 1 to 16 samples can be processed simultaneously with tens of thousands of genes profiled per sample. The Sentrix Array Matrix (SAM) contains 96 arrays, each of which us a hexagonal fibre-optic bundle with approximately 50,000 beads and 1,500 distinct bead types. Thus, a 96 samples can be examined simultaneously on a single SAM.

1.3.4 Applications of Illumina Microarrays in Systems Biology of Influenza Virus.

Expression profiling studies aims to deduce the expression level of many thousands genes simultaneously under different conditions of interest. In a simple scenario, one might try to find genes that have a distinct differential expression between normal and diseased tissue condition. In order to avoid biases of the expression data result, it is important to make the measurement for all possible genes and to have to replicate for the sample of interest. This is extremely crucial when we are investigating a system for which a little prior knowledge is available. A most popular use of microarray is in studies of disease dynamics, where one might get insights into the disease dynamics at the molecular level. Illumina expression arrays provide a good opportunity to measure a very large number of transcripts at once, giving researchers a detailed picture of gene expression in their sample of interest. The ability to make 30 observations for a specific gene of interest is of prime advantage, which might reduce some of the measurement errors inherent in microarrays. As the replicates of a specific gene of interest is spread over the array surface, minimising the effects of spatial artefacts. Moreover, the probe sequences are attached to beads rather than the array surface. Therefore, hybridisation should only take place at the bead level and the area between the beads shows no fluorescence. Hence, we would expect the background level to be low and consistent on the array surface. BeadArrays also provides reliable estimates of mean and variances across samples because of not only they provide more replicates of a particular gene on one array but their ability to process more arrays in parallel. Furthermore, by being

able to hybridise and scan multiple arrays simultaneously, we would hope to reduce the batch effects associated with running microarrays on different days.

The first step to a systems biology approach is the data assembly, which should be as extensive as possible. Investigating the molecular mechanisms of the severe influenza is important in controlling complications and reducing the pulmonary damage in the host. Comprehensive genome wide expression data involving both innate as well as adaptive immune response might help to understand the role of host's response during influenza virus infection. In recent years, many whole genome expression changes analysis has been studied in lungs of mice, and other model organisms [45, 58] after infection with different influenza viruses (including different IAV subtypes) causing mild or severe disease [1, 59–62]. All most all the studies so far have focussed on the innate immune response of the host. Until now, only one study integrated both innate and adaptive immune response, and described different phases of disease as temporal changes in gene expression profile in lungs of mice [59]. Moreover, there is no study, which correlates the gene expression profile data with the clinical outcomes or the phenotype data and to explain severe influenza with the biphasic course of disease. Recently, we used mathematical modelling to explain the biphasic course of influenza caused by the A(H1N1)pdm09 isolate Jena/5258 and suggested the second disease peak was due to high inflammation [22]. However, we failed to show the dynamics behind it due to the limited amount of data. Further experimental studies with mpJena/5258 in mice after demonstrated a fast intra-host evolution of the virus based on HA-222D/G polymorphism [2]. Viral evolution was accompanied with changes in organ tropism and different antibody recognition that could be one of the reasons for second peak of the body weight change and clinical score observed during the second lung passage of the virus [2]. Furthermore, a very strong lung histopathology was observed underlining the suggestion on the impact of inflammation.

In my thesis, I have performed a comprehensive analysis of whole genome expression changes of the lung of mice infected with the pH1N1 mpJena/5258 consisting HA-222D/G quasispecies to gain insights into pathogenesis at the transcriptional level. We studied both, the innate and adaptive host immune response in order to identify key regulatory interactions during infection. A reverse engineering strategy was applied to infer a gene regulatory network. Since the complexity of network structure increases with the number of genes, a small number of 20 differentially expressed genes (DEGs) was selected to be included in the network. In order to identify key genes, we performed functional categorization and identified significantly overrepresented functional categories.

Chapter 2

Materials and Methods

2.1 Phenomological Data

The pathogenicity of influenza viruses, A/Jena/5258/09 (Jena/5258), A/Jena/5555/09 (Jena/5555), A/Jena/2688/10 (Jena/2688) and European A/swine/Bakum/1832/00 (H1 N2) (Bakum/1832), were compared in BALB/c mice at the Jena University Hospital, department of Virology and Antiviral Therapy, Jena as described recently in Manchanda et al, [22].

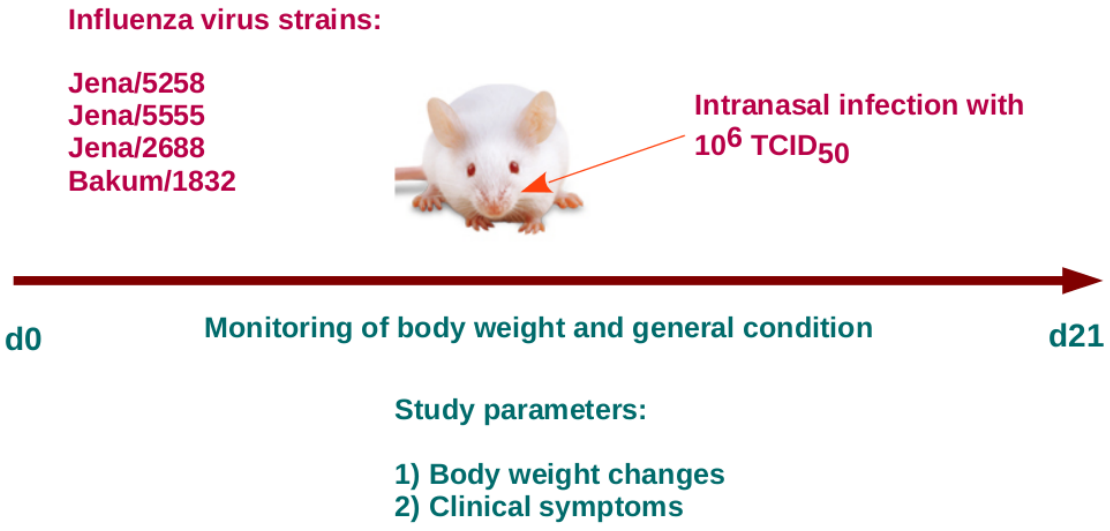


FIGURE 2.1: Pictorial representation of experimental setup. The mice were intranasally infected separately with 10^6 $TCID_{50}/20\mu l$ dose of 4 different influenza virus strains. The body weight loss change as well as clinical symptom score were observed for 21 days post infection.

Briefly, the A(H1N1)pdm09 influenza virus strains, Jena/5258, Jena/5555 and Jena/2688 were isolated in MDCK cells from respiratory specimen that originated from patients with clinical symptoms of influenza infection. The European swine H1N2 influenza virus, Bakum/1832, was obtained from a nasan swab of a diseased pig [63, 64]. Experiments were performed in female BALB/c mice, weighing 16-18g as shown in Fig. 2.1. After isoflurane anesthesia, five mice were inoculated intranasally with 10^6 $TCID_{50}/20\mu l$ of each virus in EMEM and three mice were mock infected for control.

Body weight and clinical symptoms score were monitored for 21 days after virus infection. Mice that lost more that 25% of the body weight of their initial body weight were sacrificed and the clinical score of 7 were given to them. A laboratory clinical symptom score was used to assess the severity of the disease. It ranged from 0 to 7, while scoring

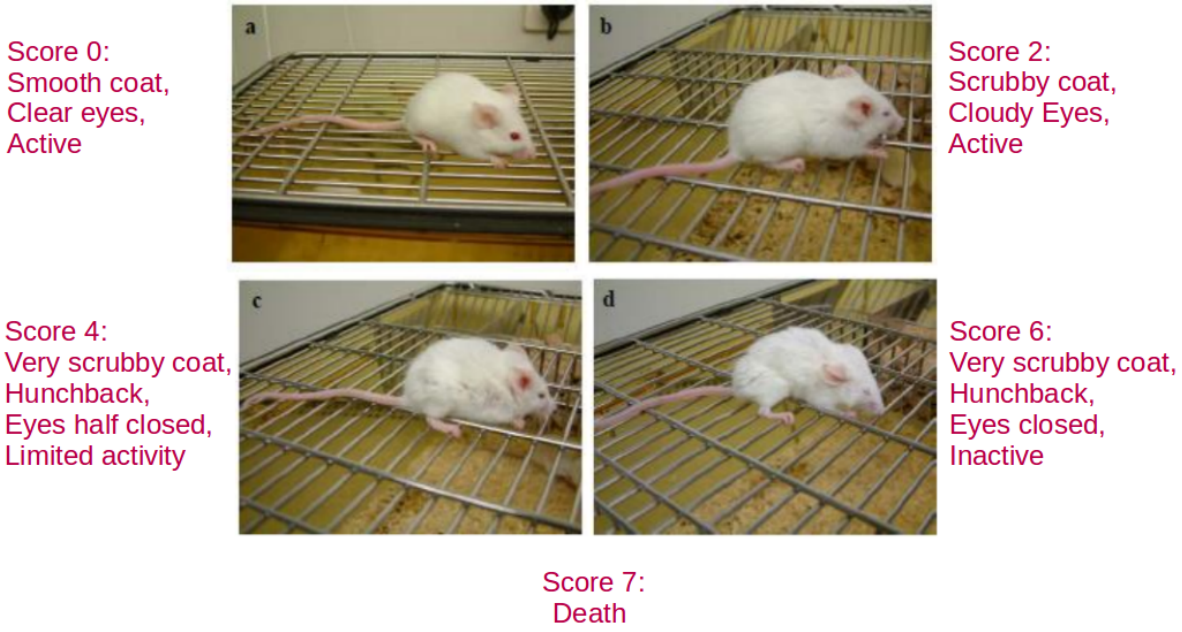


FIGURE 2.2: The general clinical scoring system, The score ranges between 0 to 7, while scoring is based on smooth coat, clear eyes and physically activity of mice. The photographs were kindly provided by the Jena University Hospital, department of Virology and Antiviral Therapy, Jena.

is based on smooth coat, clear eyes and physical activity of mice as shown in Fig. 2.2 and following:

1. scrubby coat in the neck
2. scrubby coat in the neck and on the back
3. scrubby coat on whole body, incipient hunchbacked posture

4. scrubby coat, hunchback posture, inactivity, eyes half closed
5. scrubby coat, hunchback posture, inactivity, eyes closed
6. scrubby coat, hunchback posture, completely inactive, eyes closed
7. mouse deceased or lost more than 25% of initial body weight

Interestingly, Smith and Perelson [36] suggests that using a clinical symptom score throughout the infection process could offer a new perspective into the characteristics of an infection. As symptoms score is easily obtainable and reflects how sick a host is. The mean values and standard deviation for the clinical symptom score (S) as well as Body weight changes in comparison to control were calculated and shown in Fig. 2.3 and Fig. 2.4 respectively.

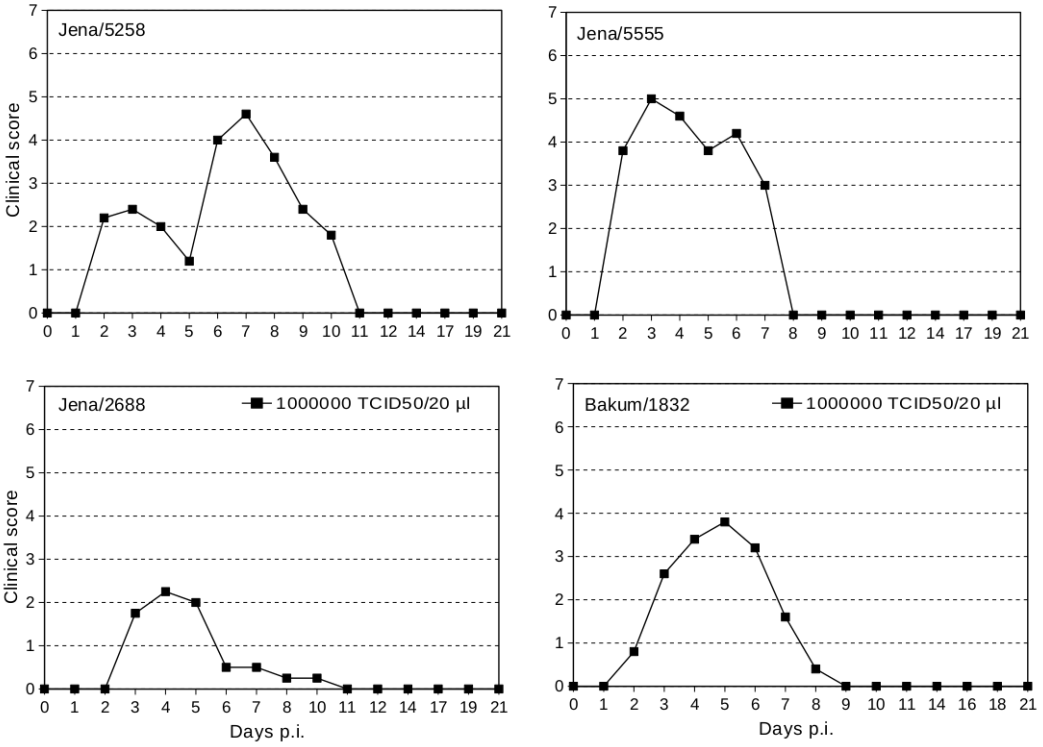


FIGURE 2.3: Clinical symptoms score for mice infected with four different influenza virus strains. Top left: A(H1N1)pdm09 Jena/5258; top right: (H1N1)pdm09 Jena/5555; bottom left: (H1N1)pdm10 Jena/2688; bottom right: European swine H1N2 Bakum/1832 (full squares).

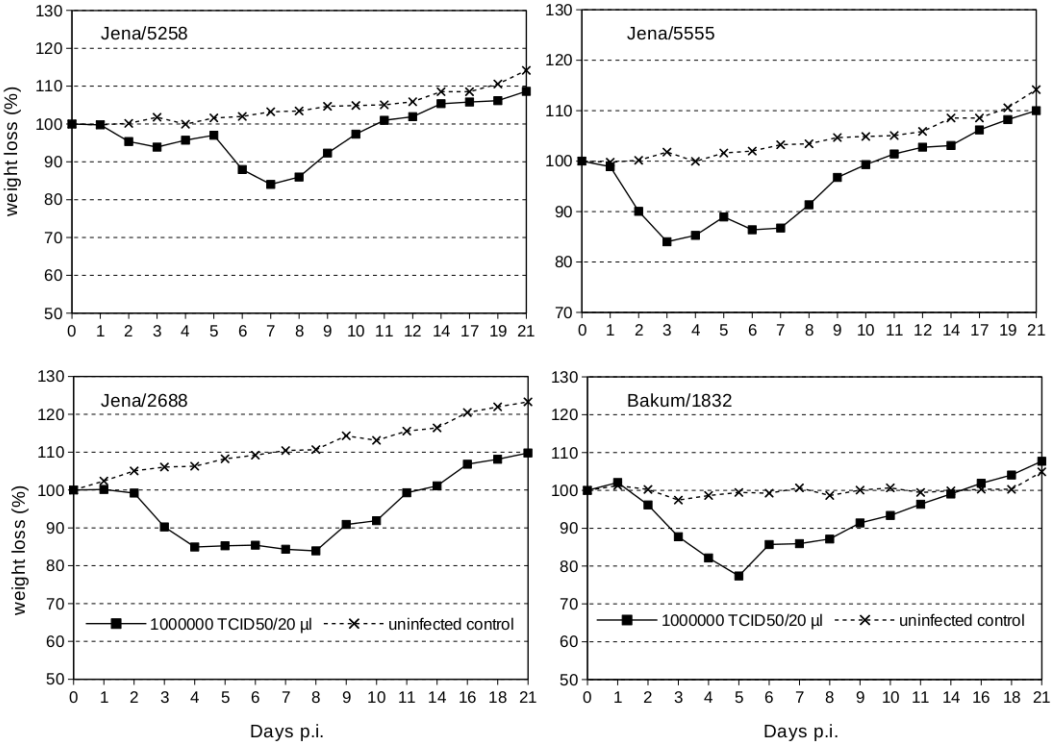


FIGURE 2.4: Body weight changes after infection with different virus strains. Top left: A(H1N1)pdm09 Jena/5258; top right: (H1N1)pdm09 Jena/5555; bottom left: (H1N1)pdm10 Jena/2688; bottom right: European swine H1N2 Bakum/1832 compared with uninfected mice. ■ represents data for virus infection whereas × represents data for uninfected control mice.

2.2 Gene Expression Microarray Data

The study design of the animal experiment was described elsewhere [2] and experiments were performed by PD Dr. Michaela's group at Institute of Virology and Antiviral Therapy (IVAT), Jena University Hospital. Briefly, 7- to 8-week-old female BALB/c mice (16 – 18 g; Charles River, Bad Sulzfeld, Germany) were infected intranasally with 106 TCID50/20 μl of the influenzavirus mpJena/5258 under isoflurane anesthesia. Mock-infected mice were used as control. Five to 10 mpJena/5258-infected mice were sacrificed on day 1-7, 9 and 12 p.i. The left lung lobe was fixed in formalin and used to evaluate lung pathology after hematoxiln-eosin staining [2]. One of the right lung lobe was frozen for RNA analysis in RNAlater stabilization solution (Ambion by Life technologies, Darmstadt, Germany).

The microarray experiments were performed in the lab of Dr. Ralf Claus at Department of Anaesthesiology and Intensive Care Medicine, University Hospital Jena. Briefly, prior to gene expression experiments total RNA integrity was confirmed using the Experion^{TM}

automated gel electrophoresis system (BioRad, Munich, Germany). cRNA sample preparation for hybridisation on the Illumina gene expression platform was performed using the TargetAmp^{TM} Nano-g^{TM} Biotin-aRNA Kit for the Illumina^{TM} System (Epicentre/Biozym, Hess. Oldendorf, Germany) starting with 250 ng of totalRNA. Samples were hybridised according to manufacturer instructions on Mouse Ref-8 v2.0 Bead Chips (Illumina, San Diego, USA). Each chip is comprising probes of 25700 coding and non coding RNA transcripts. Read outs of hybridisation signal intensities were performed on an iScan Bead Array scanner (Illumina, San Diego, USA), data pre-processing including spot detection, gene mapping and averaging of replicates was performed with iScan Control Software and GenomeStudio software (Illumina, San Diego).

2.3 Mathematical Modelling

2.3.1 Mathematical Model Notation

The mathematical model representing the within-host influenza dynamics consists of a system of three differential equations in which the dependent variables represent the virus pathogenicity (P), antiviral immune defense (D) including both the innate immune response and the adaptive immune response, and inflammation due to pro-inflammatory response (I). The mathematical equations of our reduced model are:

$$\begin{aligned}
\frac{dP}{dt} &= \alpha P\Big(1-\frac{P}{k_p}\Big) - \beta D\Big(\frac{P}{P+0.01}\Big) \\
\frac{dD}{dt} &= \gamma P - \theta D \\
\frac{dI}{dt} &= \epsilon f(D) - \rho I \\
f(D) &= 1 + \tanh\Big(\frac{D-\delta}{\omega}\Big) \\
S &= P + I
\end{aligned} \tag{2.1}$$

The virus pathogenicity P represents the virulence of the virus strains. The dynamics of change of P is described by eq. 2.1. It depends on viral infection and the immune response of the host. The first term is parameterized by the virus infection rate (parameterized by α) and the maximum primary pathogenicity (parameterized by k_P). The second term of eq. 2.1 represents the efficiency of the early immune response to the

virus (parameterized by β). For very small values of P this second term goes to zero due to the Michaelis-Menten function parameterized by a small and fixed Michaelis-Menten constant 0.01 (smaller values or other functions that go to zero if P becomes zero do not change the model behavior) to avoid that P becomes negative.

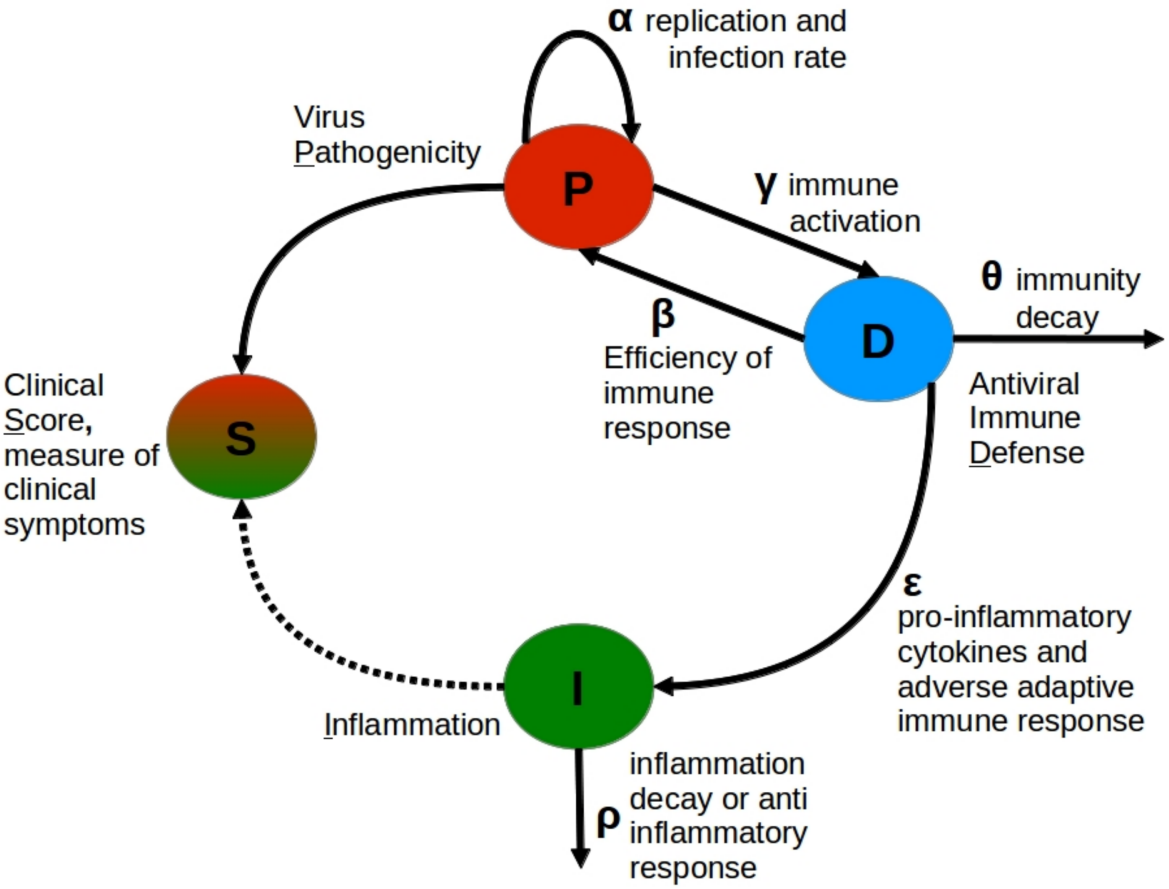


FIGURE 2.5: Pictorial representation of mathematical model developed for Influenza dynamics. The model variables P (pathogen), D (defense), I (inflammation), and S (clinical score) as well as the relations between them according to Equations 2.1

The antiviral defense is modeled by the variable D including both the innate immune response and the adaptive immune response. Although the defense system is very complex, in the small-scale model the change of the defense system is modeled in eq. 2.1 by only two terms. The first term represents the activation of the defense systems in response to the pathogen (parameterized by γ), and the second term represents the immunity decay (parameterized by θ), which means that once the pathogen has been cleared off from the host system, the defense system should come to rest. Both terms are modeled by first-order reactions.

The third model variable I represents the inflammation due to pro-inflammatory response. Its change is described by eq. 2.1. This model formulation is derived from

Table 2.1: Description of model variables and parameters.

Model Variables		
t	Time (d)	
P	Virus pathogenicity	
D	Antiviral immune defense	
I	Inflammation due to pro-inflammatory response	
S	Clinical Score (measure of clinical symptoms)	
Model Parameters	**Description**	**Value**
α	Viral infection rate	Estimated (d^{-1})
k_P	Maximum primary pathogenicity	Estimated
β	Efficiency of immune response	Fixed: 1 d^{-1}, see text 3.1.1
γ	Rate of activation of immune response	Estimated (d^{-1})
θ	Convalescence rate of immune defense	Fixed: 0.01 d^{-1}, see text 3.1.1
ϵ	Inflammation activation rate	Estimated (d^{-1})
δ	Triggering threshold values of the defense for pro-inflammation	Estimated for biphasic course only
ω	Tolerance value of the defense for the chronic inflammation	Estimated
ρ	Relaxation rate of inflammatory response or Anti-inflammatory response	Fixed: 1.82 d^{-1} [48]
P(0)	Viral Pathogenicity, initial value	Fixed: 0.01 [28]
D(0)	Antiviral immune defense, initial value	Fixed: 0.00
I(0)	Inflammation due to pro-inflammatory response, initial value	Fixed: 0.00

Kumar et al. [46], and describes a combined effect of cytokines and chemokines, and stimulatory effects of tissue damaged and dysfunction.The first term according to Kumar et al. is a hyperbolic function of antiviral defense system of the host D which strength is parameterized by ϵ. The parameter δ represents the triggering threshold value for the chronic inflammation along with the parameter ω, which represents the strength

of prolongation of the inflammation. The ratio of δ/ω represents the intensity of the inflammation which can explain whether the virus triggers the chronic inflammation or the acute one. The second term in I represents the relaxation of the inflammation (parameterized by ρ) which could be the anti inflammatory response by the host. This part of inflammation (denoted by I) does not help in removing the pathogen but leads to collateral damage to the neighboring tissue [1, 43, 65–67] contributing to the influenza virus-induced symptoms (clinical score S) as modelled by S. Fig. 2.5 displays the model structure and Table 2.1 lists the model parameters.

2.3.2 Model Simulation and Parameters Optimisation

The model formulated by the Eqs. (1)–(5) in 2.1 is numerically solved using the R-package deSolve from Soetaert [68]. The LSODA method of integration was employed using a step size of 0.01. The R-package FME from Soetaert [69] was used for the estimation of parameter values and their impact (sensitivity) on the model output S. The Levenberg–Marquardt algorithm from Moré [70] has been used to fit the model variable S to the respective data (observed clinical score) as implemented in the modFit function of the package FME. The modCost function from the package has been used in the fitting procedure and is the sum of the variable costs, which are scaled according to the number of observations. The variables cost itself is calculated as the sum of squared weighted residuals, where the standard deviation Std of the mean of the observed clinical score S is used as weight.

Finally, a Markov Chain Monte Carlo (MCMC; implemented in the package FME [69]) simulation was used to estimate the uncertainty of the parameter estimates as resulted by the model fit. The first 500 iterations were removed which is known as burn-in so as to make sure the chain was close enough to target distribution. The parameters sets that gave the highest probability to fit the model have been chosen as the final parameter sets for the model fitting. All the parameter values and variables are positive.

2.3.3 Parameters Sensitivity and Collinearity Analysis

The sensitivity and identifiability analysis were done by the functions *sensFun* and *collin* outlined in the R-package FME. Briefly, this is done by estimating at each data point the derivative of the clinical score S with respect to the parameter values p_j . The sensitivities are dimensionless, as they are scaled with respect to the variables and parameter values:

$$S_{ij} = \frac{\partial S_i p_j}{\partial p_i S_i} \tag{2.2}$$

where S_i corresponds to a data point and p_j to a parameter. These sensitivity functions allows us in determining which parameter is least important to the model output and these least sensitive parameters sets can be easily fixed from the identifiability analysis and model calibration.

Raue et al [71], distinguished between structural and practical identifiability. A structural identifiability analysis gives insight into which parameters can be simultaneously estimated, given noise free data and a model that can fit the data perfectly. A structural non-identifiability is caused by redundant parameterisation. Here, the structural identifiability was evaluated by investigating the pairwise linear dependence or collinearity of all possible parameters sets. If the 'collinearity index' calculated by the function collin exceeds a critical values typically chosen to be 10–15, then the parameter set is regarded as poorly identifiable [32].

In practice, all measurements have error and the model is non-perfect and this causes the uncertainty of the parameters values characterized by confidence regions. The practical non-identifiability is characterized by an infinitely extended confidence region in increasing and/or decreasing direction. For the estimation of confidence intervals the data-dependent probability distribution of the parameters have to be derived. Raue et al. introduced the profile likelihood to estimate the confidence intervals. In the present work, a Bayesian method, in particular the Markov Chain Monte Carlo approach using the DRAM method (Delayed Rejection Adaptive Metropolis) [72] as implemented in FME was applied to visualize the confidence regions by pairwise scatter plots and to estimate the confidence intervals.

2.3.4 Markov Chain Monte Carlo Algorithm

Bayesian inference provides a framework for explicitly accounting for modelling uncertainties, the essential characteristics of Bayesian methods being the probability distribution for describing the parameter and model uncertainty. Analysing the parameter uncertainties by using the formal likelihood function is the heart of Bayesian inference. Markov Chain Monte Carlo (MCMC) refers to an important class of Bayesian methods, designed to sample from a probability distributions by constructing a Markov chain that has the desired distribution as its equilibrium distribution. MCMC is based on the most popular mathematical property of Markov chains, as they relate to Monte Carlo sampling and distribution estimation [73]. A Markov chain is a sequence of a random variable for which the future state solely depends on the current state irrespective of the events that precedes it i.e.,
$p(\mathrm{X}_{n+1}|Xn, X_{n-1}, X_{n-2}..., X_1) = p(X_{n+1}|X_n)$.

Markov chains can be used to generate samples of the posterior distribution of the model parameters, using a random walk approach [74]. Many MCMC algorithms have been developed with the aim of constructing and developing statistically more relevant Markov chain, some of them have been explained below.

2.3.4.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is one of the simplest and most successful MCMC method for obtaining a sequence of random samples from any probability distribution for which a direct sampling would be difficult. The key idea behind this algorithm is to randomly generate perturbations of the current state, x_i, of the Markov chain and either accept or reject the new perturbed state, x_j, depending on how the probability of the new state changes is relative to the current state. The full MH- algorithm takes up the following form:

Suppose we have sampled the states as: $x_i, ..., x_n$.

set initial state to x_i and do :

Step(1) Sample from the proposal distribution $q(x_j|x_i)$

Step(2) Evaluate $\alpha = \dfrac{q(x_i|x_j)p(x_i)}{q(x_j|x_i)p(x_j)}$

Step(3) Accept the proposal state, x_j with probability $\min[1, \alpha]$, otherwise keep the current state x_i as the new state.

(2.3)

where $Q = q(x_j|x_i)$ (simply q_{ij}) represents the proposal density, sometimes also called as jumping distribution, posing the probability of selecting state x_j while being in state x_i as well as an acceptance distribution $R = (r_{ji})$ with r_{ji} being the probability of accepting the state s_j while being in the state s_i [75]. For a Metropolis algorithm, it is essential that the proposal distribution Q is symmetric, i.e., $q_{ij} = qji$. Starting from a state x_i at time t, the Markov chain generates a new state x_j, which totally depends on the previous state x_i, with a probability q_{ji}. This could for example be done with a (multivariate) Gaussian function centered on the current state $x_i : x_{ji}N(x_i, \sigma^2 I)$. This proposal is accepted as the next value if α drawn form the uniform distribution U (0, 1) satisfies:

$$\alpha < \frac{P(x_i)q_{ji}}{P(x_j)q_{ij}}$$

If the new state is accepted then the procedure is repeated, proposing a new state, x_{j+1}, otherwise the new state is rejected and the new state will be the current state, x_i.

The **adaptive Metropolis** (AM) algorithm is based on the classical random walk Metropolis algorithm [76] and its modification. The key distinction of the AM algorithm, from MH algorithms, is its continuous adaptation towards the target distribution via its calculation of the covariance of the proposal distribution using the history of chain. Keeping this characteristic intact, the proposal distribution is updated using the information gained from the posterior distribution thus far.

An important advantage of the AM algorithms is that it starts using the accumulating information right from the beginning of the simulation, which ensures that the search becomes more effective even at an early stage of the simulation. To be more precise, let us assume that at time t, the Markov chain is $x_i, ..., x_n$. The new proposal distribution for the new state is then a Gaussian distribution with a mean at the current state x_i and covariance given by $s_d R$, where R is the covariance matrix determined by the spatial distribution of the states $x_i, ..., x_n \in \mathbb{R}$. The scaling parameter s_d depends only on the dimension d of the vectors, to ensure reasonable acceptance rates of the proposed states. This adaptation strategy allows the proposal distribution to reach an appropriately scaled Gaussian approximation of target distribution, which increases the efficiency of the simulation.

In AM, the proposal distribution is continuously adapted by setting:

$$C_j = \begin{cases} C_i & i \leq j \\ s_d.cov(x_i, ..., x_n) + s_d \epsilon I_d & i > j \end{cases}$$

where $\epsilon > 0$ is a constant value which is added on the diagonal of the covariance matrix so as to prevent it from becoming singular and I_d represents a $d-$dimensional identity matrix. Initial covariance of C_0, an arbitrarily positive value, according to the best prior knowledge can be used to start the algorithm.

Delayed rejection (DR) is a way of modifying the standard MH algorithm [73], to improve the efficiency of the standard algorithm. The basic idea behind DR is that, upon rejection in a MH for the new state, x_j, instead of going back to the initial state, x_i, a second new state, x_{j+1}, is proposed which is not only depends on the initial state but also on the rejected state. The acceptance probability of the second state is computed

so as to retain the reversibility of the Markov chain with respect to the proposal distribution. As seen in 2.3, the acceptance probability of the new state in standard MH is :

$$\alpha(x_i, x_j) = 1 \wedge \frac{q(x_i|x_j)p(x_i)}{q(x_j|x_i)p(x_j)}$$

However, in DR, upon rejection a new state x_{j+1} is proposed with an acceptance probability depending not only on the initial state x_i but also on the state we have proposed and rejected, x_j. The second state proposal is accepted with probability:

$$\alpha(x_i, x_j, x_{j+1}) = 1 \wedge \frac{p(x_{j+1})q(x_{j+1}, x_j)q(x_{j+1}, x_j, x_i)[1 - \alpha(x_{j+1}, x_j)]}{p(x_i)q(x_i, x_j)q(x_i, x_j, x_{j+1})[1 - \alpha(x_i, x_j)]}$$

2.4 Microarray Data Analysis

As already explained before that I have used Illumina microarrays for analysing genome expression data, I will now describe in brief about the algorithms utilized by Illumina in R package lumi [77]. The basic preprocessing steps, such as the image analysis and background correction methods, to be applied to almost all types of Illumina data, however, the methods I describe for Illumina microarray are specific to analysis of expression data.

2.4.1 Image Processing, and Background Correction

The foreground estimation algorithm used by Illumina is a two step process described in detail by Kuhn et al., [56]. In summary, these two steps are following:

1. All pixel intensities are altered using a sharpening transformation, which mean, the intensity of a particular pixel is either made higher or lower if its intensity is higher or lower in corresponds to neighbouring surrounding pixels.

2. Foreground intensities are calculated as a weighted average of signals obtained using the four pixels nearest to each bead centre. Sharpened pixels intensities are used in calculation and the value returned is unlogged.

Background intensities are calculated using an average of the five lowest unsharpened pixels around each bead centre covering the area of 17×17 pixel. Background corrected intensities are then calculated by subtracting the background estimates from the foreground estimates. The function *lumiB* was provided by *lumi* for Background correction.

2.4.2 Normalisation

The *lumi* package provides several normalisation options. It includes a variance-stabilising transformation (VST) [78], that takes advantages of the technical replicates available on every Illumina microarray. Between chip normalization was done by Variance Stabilisation and Normalisation (*vsn*) method using *LumiN* function with aim that the intensities should become independent of the mean. The *vsn* is used which builds upon the facts that the variance of microarray data depends on the signal intensity and that a transformation can be found after which the variance is approximately constant. It is like the logarithm at the upper end of the intensity scale, approximately linear at the lower end, and smoothly interpolates in between. The position of the cross-over point and the slope of the linear part depends on the error distribution of the data. It also incorporates the estimation of normalisation parameters. The *vsn* assumes that less than half of the genes on the arrays are differentially transcribed across the experiment. An advantage of vsn-transformation over $log-$transformation is that *vsn* works also on values that are negative after background correction. The following command will do the *vsn* normalisation.

$Norm\_data = LumiN(Unnorm\_Data, method = "vsn")$

2.4.3 Identification of Differential Expressed Genes (DEG's)

The R-package *Limma* was used to identify the differentially expressed genes. Limma provides functions such as *topTable()* and *decideTests()*, which summarises the results of the linear model, perform hypothesis tests and adjust the p-values. The basic statistics used for significance analysis id the moderated t-statistics, which is computed for each probe and for each contrast. This has the same interpretation as an ordinary *t-statistics* except that the standard errors have been moderated across genes, which means it shrunk towards a common value, using a simple Bayesian model [79]. Moderated t-statistics lead to p-values in the same way that ordinary t-statistics do except that the degrees of freedom are increased, reflecting greater reliability associated with the smoothed standard errors. A number of summary statistics are presented by topTable() for the top genes and the selected contrast. The M-value (M) for instance is the value of the contrast (for e.g. diseased vs control). Usually this represents a log_2-fold change between two or more experimental conditions, sometimes it represents a log_2-expression level too. The Avalues (A) is the average log_2-expression level for that gene across all the arrays and channels in the experiment. Column **t** is the moderated t- statistics. Column **p-value** is the associated p- value after adjustment for multiple testing. The most popular form of adjustment if **fdr** which is Benjamini and Hochberg's method to

control the false discovery rate [80]. The adjusted fdr p-value means, if all the genes with p-value below a threshold, e.g. 0.05, are selected as differentially expressed then the expected proportion of false discoveries in the selected group is controlled to be less than the threshold value, in this case 5%. The B-statistics (B) is the log-odds that the gene is differentially expressed. For example, suppose a gene X has B value of 1.5, this means that the odds of differential expression is $\exp^{1.5}$, which is equals to 4.48. The probability that the gene X is differentially expressed is $(\frac{4.48}{1+4.48}) = 0.82$, i.e., the probability is about 82% that this gene is differentially expressed. The B- is automatically adjusted for multiple testing by assuming that 1% of genes are expected to differentially expressed with respect to the control.

The **ebayes()** function also computes one more vital statitstics. The moderated F-statistics (F) combines the t-statistics for all the contrasts into an overall test of significance for that gene. The F-statistic test whether any of the contrasts are non-zero for that gene, i.e., whether that gene is differentially expressed on any contrast.

2.4.4 Cluster Analysis, and Gene Regulatory Network Modelling

Cluster analysis is a collection of statistical methods, which aims at identifying groups of samples (in our case genes) that behaves similarly or show similar characteristics. In general, it is also called as look-a-like groups. The easiest way is to partition the samples using the measurement that captures similarity or distance between samples. In this way, clusters and groups are interchangeable words. Typically, in clustering methods, all the samples with in a cluster is considered to be equally belonging to the cluster.

The clustering algorithms are classified into two broad categories namely, hierarchical and non-hierarchical algorithms. In the hierarchical, one construct a hierarchy or tree like structures of the entities involved (genes in our case). While, in non-hierarchical method, one tends to find a position in the measurement that takes as the central place and distance is measured from that particular central point. Identifying a right central position is quite a challenging task and hence non-hierarchical methods are less popular but sometimes effective when datasets contains many hundred entities. In my thesis, I will focus on one of the non-hierarchical algorithm known as fuzzy c-mean clustering algorithm.

2.4.4.1 Hierarchical Clustering

Hierarchical clustering algorithms are of two types: agglomerative and divisive clustering.

Agglomerative clustering is based on the union between the two nearest clusters (or entities). The beginning condition is realized by setting each entity as a cluster. After a few iterations it reaches the final clusters. Basically, it is a bottom-up approach. Whereas Divisive clustering is a top-bottom approach where one starts from one cluster containing all the data entities. At each steps, clusters are successively split into smaller clusters according to some dissimilarity.

Given a set of N objects, $S = \{S_1, S_2, ...S_N\}$ to be clustered and a function of distance $D(c_i, cj)$ between two clusters or entity c_i and c_j, build a hierarchy tree on $S : c_i, c_j \subset S, c_i \cap c_j$. The basic hierarchical algorithm has following steps:

1. Start by assigning each entity to a cluster, $c_i = s_i(i = 1, ...N)$, so that if you have N objects, you will have N clusters $\ell = \{c_1, c_2, ...c_N\}$, each containing just one entity.

2. Find the pair of cluster, (c_i, c_j) such that $D(c_i, c_j)) \leq D(c_{i'}, c_{j'}) \forall c_{i'} \neq c_{j'} \in \ell$ and merge them into a single cluster $c_x = c_i \cup c_j$. Therefore, now we have $\ell - 1$ number of new cluster, as we have merged the clusters $c_i, c_j) into c_k$, so that now we have one cluster less.

3. Compute distance (similarities) between the new cluster and each of the old cluster.

4. Repeat steps 2 and 3 until all entities are clustered into a single cluster of size N.

Step 3 in the hierarchical clustering algorithms can be done in different ways, which is what distinguishes $single - linkage$ from $complete - linkage$ to $average - linkage$ clustering. All of these differ in the definition of distance and what defines largest distance as statistically no-distance or zero-distance. Most of the time, the distance is based on the *Euclidean* distance in the sample axes. The main weakness of hierarchical clustering methods include that they do not scale well.

2.4.4.2 Non-Hierarchical Clustering

One of the non-hierarchical clustering methods is the partitioning method. Suppose, x as a given number of clusters as the objective and the partition of the entities to obtain the required x clusters. In contrast to the hierarchical clustering methods, this

partitioning method permits entities to change group membership through the cluster formation process. The partitioning methods usually begins with an initial solution, after which reallocation occurs according to some optimality criterion.

k-means Clustering is one of the simplest unsupervised learning algorothms that solves the clustering problem [81]. The objective is to classify a gicen data set $S = \{S_1, S_2, ...S_N\}$ into a certain (assume initial clusters) number of clusters. The idea is to defive intial centroids, one for each cluster $c_i(i = 1, ...k)$. The basic steps involved in the algorithm is as follow:

1. Initial clusters: $\ell^* = \{c_1^*, c_2^*, ...c_k^*\}$; the intial centroids should be placed as far as possible from each other.
2. Calculate the centroids of the clusters: $u_j^i = \frac{1}{\mid c_j^i \mid} \sum_{x \in c_j^i} x$, where $j = 1, 2...k$ and i denotes the i^{th} iteration.
3. Take each point belonging to a give data set and associate it to the nearest centroid:
   $c_j^{i+1} = \{x \mid d(x, u_j^i \leq d(x, u_{j'}^i) \leq j^{'} \leq k\}$
   $\ell^{i+1} = \{c_j^{i+1} \mid 1 \leq j \leq k\}$
4. Repeat steps 2 and 3 until no more changes can be made to the clusters, that is, $\ell^{i+1} = \ell^i$. In other words centroids do not move any more.

Finally, this algorithm aims at minimising an objective function, in this case a squared error function:

$$J = \sum_{j=1}^{k} \sum_{x \in c_j} \mid\mid x - u_j \mid\mid^2$$

Although it can be proved that the k-means algorithm will always terminate, the algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. It might get stuck in a local minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. To get out of the local minimum, the k-means algorithm can be run multiple times from different initial clustering or sometime fuzzy c-means clustering technique could be used.

Fuzzy c-means (FCM is a method of clustering which allows an entity of dataset to belong to two or more clusters with a membership value. This method was developed by Dunn in 1973 [82] and later on improved by Bezdeck in 1981 [83]. It is frequently

used in pattern recognition. In simple, this algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of their distance to the cluster. Smaller the distance to the cluster center, more is its membership towards that particular cluster center. The overall algorithm is quite similar to as that of $k-means$ clustering algorithm, except the following changes:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{k} \mu_{ij}^{m} \mid\mid x - u_j \mid\mid^2, 1 \leq m < \infty$$

where m is the Fuzzines exponent and is any real number greater than 1, N is the number of data, k is the number of clusters, μ_{ij} is the degree of membership of x_i in the cluster j.

Fuzzy partitioning is carried out by an iterative optimization of the objective function as explained above, with the update of membership μ_{ij} and the cluster centre u_j by:

$$\mu_{ij} = \frac{1}{\sum_{z=1}^{k} \left(\frac{\mid\mid x_i - u_j \mid\mid}{\mid\mid x_i - u_z \mid\mid}\right)^{\frac{2}{m-1}}}$$

$$= \frac{1}{\left(\frac{\mid\mid x_i - u_j \mid\mid}{\mid\mid x_i - u_1 \mid\mid}\right)^{\frac{2}{m-1}} + \left(\frac{\mid\mid x_i - u_j \mid\mid}{\mid\mid x_i - u_2 \mid\mid}\right)^{\frac{2}{m-1}} + ... + \left(\frac{\mid\mid x_i - u_j \mid\mid}{\mid\mid x_i - u_z \mid\mid}\right)^{\frac{2}{m-1}}}$$

where $\mid\mid x_i - u_j \mid\mid$ is the distance from point i to current cluster centre j, and $\mid\mid x_i - u_z \mid\mid$ is the distance from point i to other cluster centre z.

$$u_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}}$$

Iteration stops when $\max_{ij}\left\{ \mid \mu_{ij}^{z+1} - \mu_{ij}^{z} \mid \right\} < \epsilon$, where ϵ is a termination criterion between 0 and 1, whereas z are the iteration steps. This procedure converges to a local minimum or a saddle point J.

The advantages of fuzzy c-means are that it gives best results for overlapped data set and comparatively better than the k-means algorithm. Unlike k-means where data point must exclusively belongs to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belongs to more than one cluster center.

2.4.4.3 Gene Regulatory Network Reconstruction

Constructing regulatory network from high throughput data is a key step in systems biology. The Network inference was performed as previously described using the R-package, *NetGenerator* tool [84–87].

Briefly, it is based on a set of linear differential equations and models the temporal change of expression intensity of the gene xi(t) at time t as the weighted sum of all the other n genes intensity and two external stimuli u1(t) and u2(t):

$$\mathbf{x_i(t)} = \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{w_{i,j} x_j(t)} + \sum_{\mathbf{k=1}}^{\mathbf{2}} \mathbf{b_{i.k} u_k(t)} \qquad \textbf{for i} = \mathbf{1...n}$$

The NetGenerator tool calculates the gene regulatory interaction matrix W and the components $b_{i,k}$. The parameter $w_{i,j}$, which is a component of W, represents an influence of gene j on the expression of gene i, while the parameter $b_{i,k}$ represents the impact of external stimulus given by the function $u_k(t)$. The external stimuli (input perturbation) $u_1(t)$ and $u_2(t)$ are modelled from the viral titre data for two quasi species of pIAV strains taken from Siedel et al [2], as shown in Fig. 3.8 . A non-zero weight $w_{i,j}$ defines an interaction (edge) of the inferred network where a positive weight is interpreted as activation and a negative one represents the repression or inhibition.

2.4.4.4 Linear Regression Model for Association Between Genotype and Phenotype

A multiple linear regression model was developed with a aim of finding an association between the observed phenotype data, which was quantified by the symptom score S from Manchanda et al [22], and the gene expression data. The linear regression was performed by ”lm” function in R [88].

Chapter 3

Results

3.1 Small-scale Mathematical Model for Influenza Viruses Quantification

3.1.1 Parameters Estimation and optimisation

Parameters estimation is one of the major challenging task in mathematical modelling. For estimating the parameters and their uncertainties we used R package FME and followed the procedure outlined in Soetaert et al., [69] as also explained in Method section 2.3.2.

The parameter β was eliminated by the following model transformation: $\beta^{'} = 1d^{-1}$, $D^{'} = D^{*}\beta$, $\gamma^{'} = \gamma^{*}\beta$, $\delta^{'} = \delta^{*}\beta$, $\omega^{'} = \omega^{*}\beta$. Also, The parameters ϵ, i.e., the rate at which inflammation gets activated, as well as ω and δ were only identified for Jena/5258 that exhibits a biphasic course of infection. For the other three strains the kinetics of inflammation is non-observable based on the measured clinical score S and the model Eqs. 2.1. Therefore, the parameter ϵ was set to zero and the parameter values of ω and δ were denoted as unknown (NaN) (See Table 3.1) for the infection experiments using the virus strains Jena/5555, Jena/2688, and Bakum/1832.

In the following text, the transformed parameters are used for better readability without the prime. Then, the mathematical model includes 8 parameters, four of them (α, γ, k_p, ϵ) were virus strain specific and therefore calibrated individually by fitting to the data of the experiments with different virus strains. The two parameters δ and ω are only relevant for biphasic course of infection and, thus, estimated only for the virus Jena/5258. The remaining two parameters ρ and θ are host-related and, therefore, fixed, i.e., identical for the four virus strains. For the inflammation decay we use the value

TABLE 3.1: Model parameter values. Parameter values and their 95% confidence intervals (in brackets) identified by model fit to the observed clinical scores. The parameter values NaN denote that they could not be identified. Units of the parameters are displayed in Table 2.1.

Parameter	Infection experiments			
	Jena/5258	Jena/5555	Jena2688	Bakum/1832
α	3.63	3.64	2.12	2.22
	[3.62, 5.78]	[3.31, 4.16]	[2.03, 2.29]	[2.09, 2.40]
k_P	3.23	5.69	3.72	4.85
	[2.31, 3.29]	[5.13, 6.17]	[3.02, 4.54]	[4.27, 5.36]
$\gamma^{'}$	0.51	0.28	0.52	0.29
	[0.25, 0.64]	[0.28, 0.31]	[0.44, 0.65]	[0.24, 0.32]
θ	0.01	0.01	0.01	0.01
ϵ	6.81	0.00	0.00	0.00
$\delta^{'}$	4.27	NaN	NaN	NaN
$\omega^{'}$	0.13	NaN	NaN	NaN
ρ	1.82	1.82	1.82	1.82

1.82 d^{-1} published by Canini and Carrat [48]. The fixed parameter θ were calculated by model fit to all the data sets providing a rich data set of 72 data points from four different virus strains. The values of the fixed parameters are shown in Table 2.1 and the remaining parameters were estimated.

First, the parameters sensitivity analysis was done to exclude any parameters to which the model is insensitive. Then, identifiability analysis was used to determine the minimal parameters sets for which parameters estimation is possible, given the observed data. The parameters sensitivity and identifiability analysis methods are explained in section 2.3.3. The next step was calibration, where the Levenberg-Marquardt algorithm has been used to fit the model, and the best fit values were estimated, using *modFit* function from FME package, for a selected set of parameters. Finally, sensitivity, identifiability and calibration methods were repeated iteratively. Fig. 3.1 displays the results of model fit to four experimental viral infections with four different virus strains. Table 3.1 depicts the identified values of the 8 model parameters (i.e., all but the eliminated parameter β) together with their 95% confidence intervals (CI). Finally, a Markov Chain Monte Carlo (MCMC) simulation was used to estimate the data dependent parameters uncertainty and how that affects the model output.

3.1.2 Parameters Sensitivity and Identifiability

The sensitivity analysis were done by the functions *sensFun* outlined in the R-package FME and also explained in the method section 2.3.3. The mean sensitivity values (averaged over time) of the model parameters calculated by Eq. 2.2 for the four virus strains

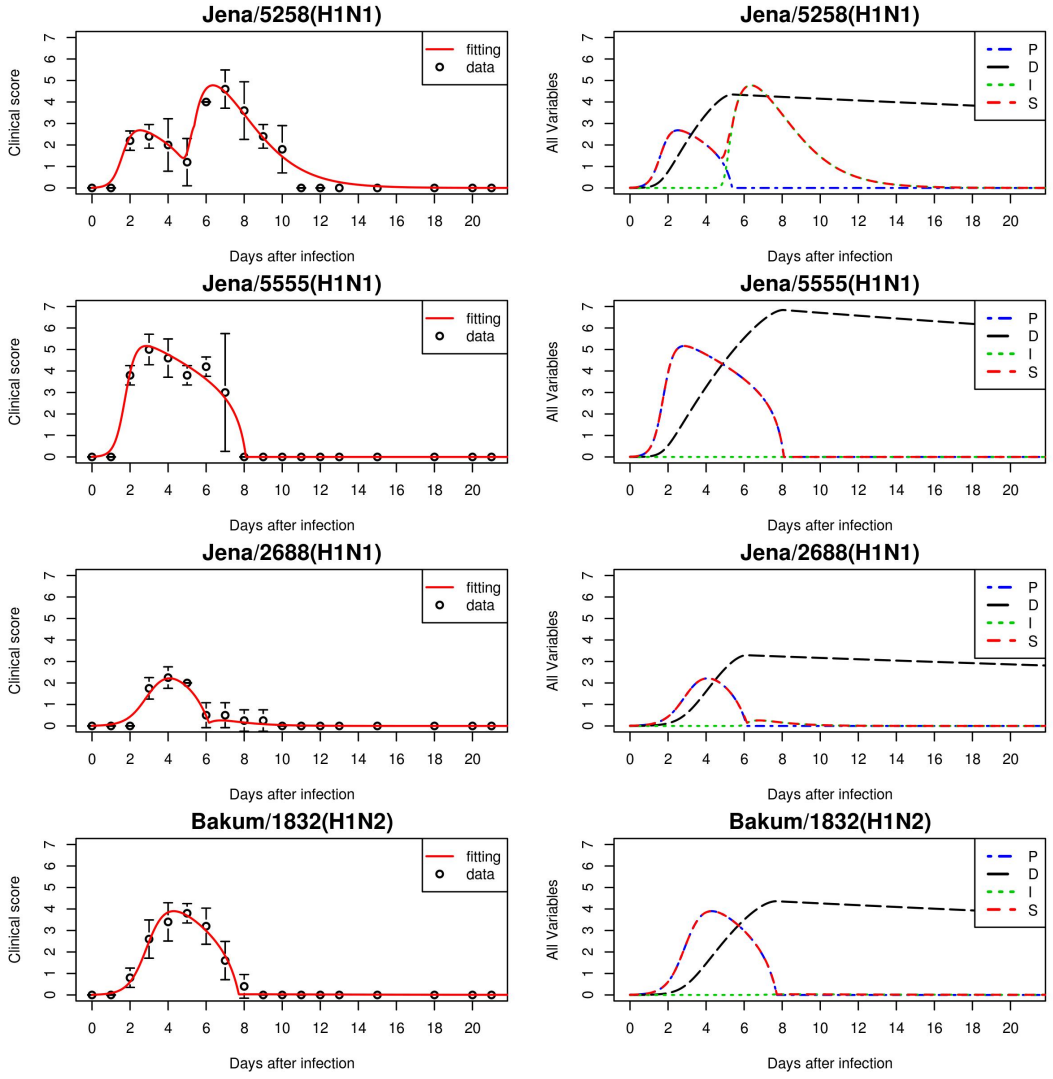


FIGURE 3.1: Model fit to the clinical score after infection with four different virus strains. Top: A(H1N1)pdm09 Jena/5258; medium: A(H1N1)pdm09 Jena/5555 and A(H1N1)pdm10 Jena/2688; bottom: European swine H1N2 Bakum/1832; all infected with 106 TCID50 /mice; black circles: observed and averaged clinical score S; bars: 'Std' (standard deviation); red lines: model kinetics S as simulated using Eqs. (1)–(5); blue dashed lines: model variable P; black dashed lines: model kinetics D ; green dotted lines: model kinetics I.(For interpretation of the references to color in text, the reader is referred to the web version of this article.)

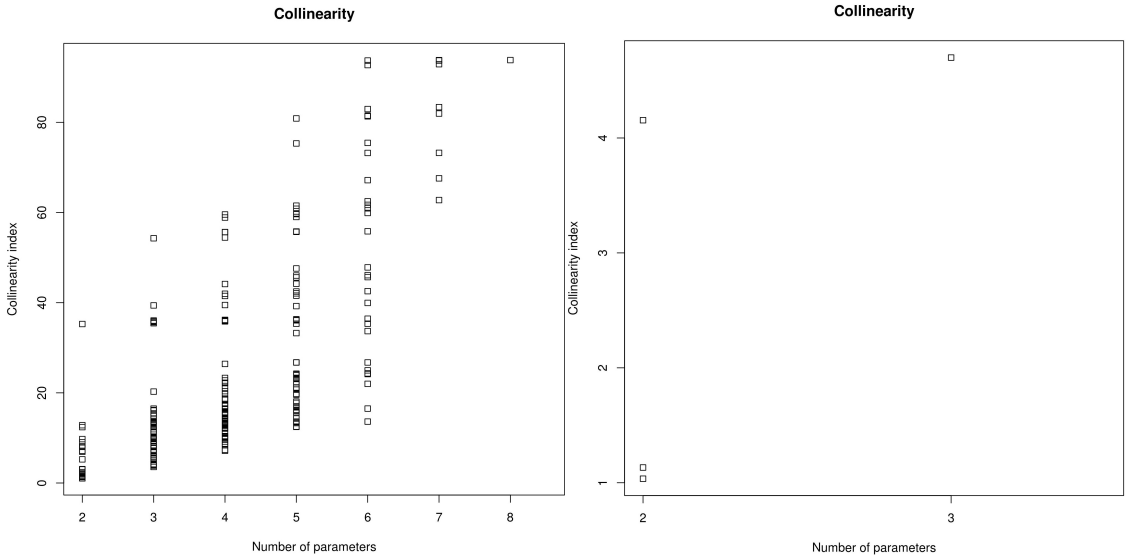
are shown in Table 3.2. As displayed in Table 3.2, we used different time intervals for averaging to exclude periods with values of S close to zero as S is in the denominator in Eq. 2.2.

To evaluate the structural identifiability of the non-fixed parameters, the collinearity index was estimated using the function *collin8* of the R-package FME. The collinearity indices for all the combinations of the 8 model parameters, i.e. α, k_P, γ, θ, ϵ, δ, ω, *and* ρ were quite high (See Fig. 3.2) and, this, demonstrate the structural non-identifiability

TABLE 3.2: Sensitivities of the model parameters. Mean value (averaged over time) of the sensitivity with respect to the read out value S.

Parameter	Infection experiments			
	Jena/5258	Jena/5555	Jena2688	Bakum/1832
α	2.88	0.99	1.78	1.67
k_P	4.57	0.39	0.46	0.46
$\gamma^{'}$	0.96	-0.12	-0.34	-0.16
θ	0.38	0.00	0.01	0.01
ϵ	0.20	0.00	0.00	0.00
$\delta^{'}$	-8.03	0.00	0.00	0.00
$\omega^{'}$	0.00	0.00	0.00	0.00
ρ	-0.46	0.00	0.00	0.00

of the whole parameter set. The collinearity indices for the combinations of three most important model parameters, i.e. α, k_P, *and* γ are shown in the Fig. 3.2b. These values are below 5. As a rule of thumb, a collinearity index less than 10–15 estimated by the function *collin* of the R-package FME is assumed to be structural identifiable [32].



(A) (B)

FIGURE 3.2: Collinearity indices for combinations of model parameters, (A) shows collinearity indices value for all combination of 5 parameters, i.e., $\alpha, k_P, \gamma, \delta,$ *and* ω, and (B) for α, k_P, γ.

Fig. 3.3 show the scatter-plot matrices visualizing the model parameter values from MCMC-generated samples, the probability distribution of the model parameters values and their pair-wise correlation coefficients as calculated by the function pairs of the package FME. For all four experiments the distributions of the parameters α, k_P, and γ

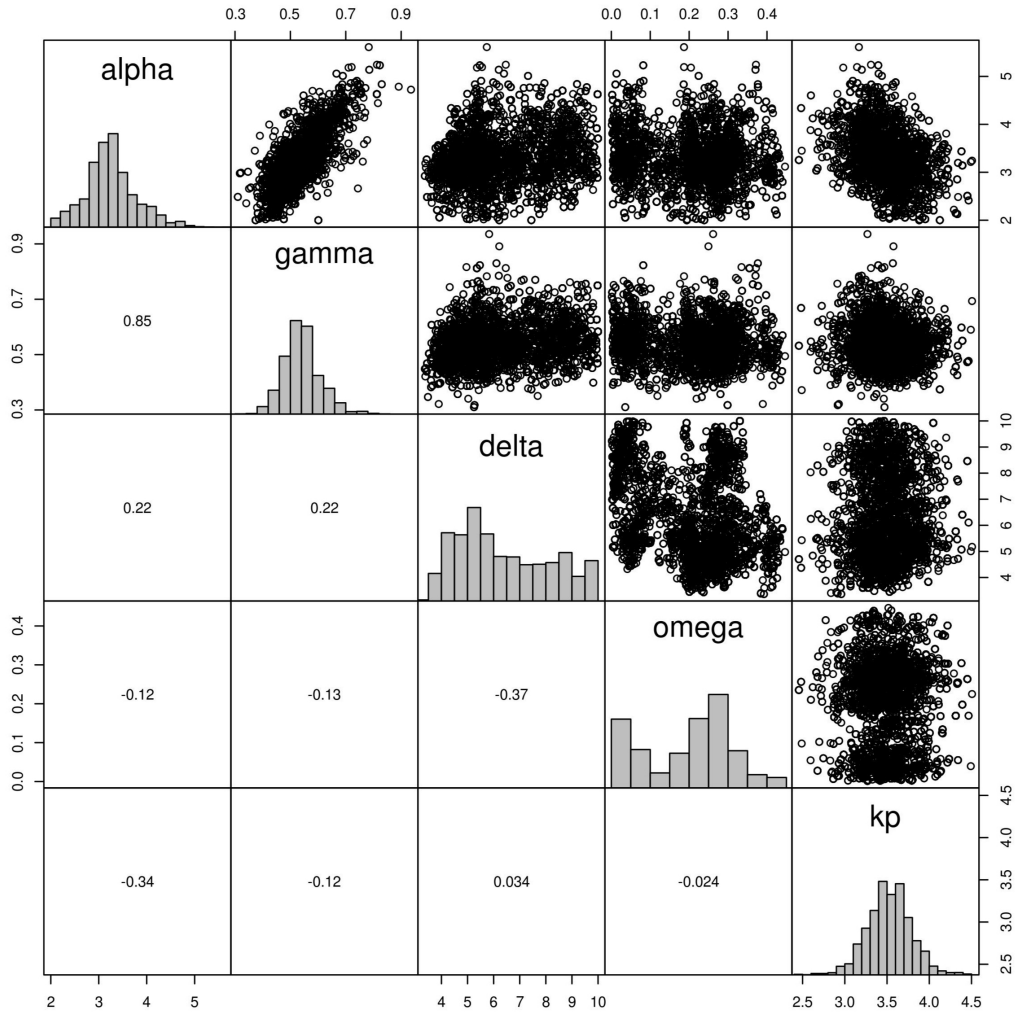


FIGURE 3.3: Scatter-plot of the Markov Chain Monte Carlo (MCMC) samples for the virus strain Jena/5258. The top right panels show the pair-wise plot of the parameters α, γ, ω, and k_P. The range of the parameters is shown at the outer axis; 1000 random parameter sets around the parameter set calibrated for Jena/5258. The probability distribution for each parameter is shown in the diagonal panels. The bottom panels give the correlation coefficients for each parameter pair.

tend to zero for both increasing as well as decreasing values and the correlation between the parameters values is low,i.e., the highest absolute value of correlation coefficients is 0.86. That indicates, that the confidence regions are finite and these three parameters are obviously practically identifiable based on the measured data of all the four experiments studied. The remaining parameters are non-identifiable at all and, therefore, they were fixed in our study. The parameter ϵ, δ, and ω were estimated as showed in Table 2.1 for the infection with virus strain Jena/5258 that shows biphasic course. The distribution of these parameters do not show finite borders as seen in Fig 3.3. For other three virus strains with a monophasic course, the parameter ϵ were set to zero and the values of

the parameters δ and ω were irrelevant as inflammation does not cause a separate peak in the profile of the clinical score S and, therefore, it is non-observable based on these easily attainable experimental data.

3.1.3 Identification of important parameters for virus quantification

The model parameters k_P, α, and γ were found to be the most discriminative and identifiable parameters that quantitatively characterise the four virus strains under investigated in this thesis. Each of the four virus strains is characterized by a specific pattern with respect to these three parameters as displayed on Table 3.3. For Jena/5258, the threshold parameter δ is the most sensitive model parameter. Here, the immune defense (D) exceeds the threshold parameter δ and triggers the onset of inflammation due to the fact that three parameters are high: (i) α, i.e., the viral replication and infection rate, (ii) γ, i.e. the rate of early activation of the immune system, and (iii) the parameter ϵ, i.e., the rate at which inflammation gets activated.

For Jena/5258, the parameter k_P called maximum primary pathogenicity is the second most sensitive parameter after the threshold parameter δ. The infection with this strain is characterized by the highest clinical score value that is caused not by a high primary pathogenicity k_P but by the inflammation ('secondary pathogenicity'). This parameter k_P has the smallest value of 3.23 for Jena/5258 and the highest value of 5.69 for Jena/5555. The confidence intervals (CI) of k_P for these two virus strains with high temporal maximum clinical score are disjunctive, i.e., without overlap. Thus, the causes of high clinical score in both infection courses are different: Jena/5555 causes the high clinical score by a high primary pathogenicity, while Jena/5258 by high inflammation. The parameter α called viral infection rate is the next most sensitive parameter for Jena/5258 with biphasic course of infection, but the most sensitive parameter for the other three virus strains with monophasic profile (See Table 2.2). It has the highest identified values for the H1N1 strains Jena/5258 and Jena/5555 that are characterized by the utmost clinical score. The CI of α for these two virus strains are disjunctive to the respective intervals of the other two virus strains Jena/2688 and Bakum/1832.

TABLE 3.3: Virus-specific pattern with respect to three model parameters. Labels 'low' versus 'high' represent disjunctive confidence intervals of the respective parameter values.

Parameter	Infection experiments			
	Jena/5258	Jena/5555	Jena2688	Bakum/1832
α	Low	High	Medium	Medium
k_P	High	High	Low	Low
$\gamma^{'}$	Medium	Low	High	Low

For all four virus strains, the parameter γ, which quantifies the rate of activation of the immune system, is the third (fourth for Jena/5258) most sensitive parameter − after k_P and α (and after δ that is only relevant for the strain Jena/5258).

In Table 3.1, the confidence intervals are only shown for the parameters α, γ and k_P. The distribution and the pairwise scatter plot of the MCMC-generated parameter values shown in Figs. 3.2a and 3.2b gives a hint to the practical identifiability of these three most important model parameters.

3.2 Reconstruction of Gene Regulatory Network from Gene Expression Data

3.2.1 Normalisation

The raw microarray data was analyzed using the "Lumi" [77] and "Limma" [89] package of the statistical language "R" [90]. Between-chip normalization was done using VSN-transformation with "lumiN" so that the distribution of intensities should become independent of the mean. See methods section for more details.

3.2.2 Identification of Differentially Expressed Genes (DEG's)

One of the basic and far most interesting goals in the analysis of microarray gene expression data is the identification of differentially expressed genes (DEGs) in comparison with the mock control. In order to control the biological and technical (experimental) variability of the experiments, one has to make an adequate number of replicates, so that statistical inference can have a more impact. Depending on the type of study, one can distinguish between the cases in which one aims to study a particular cell population or to study the disease progression. In the former case, independent replication on the level of mRNA isolation can be obtained, whereas in the latter case, one can obtained samples/replicated from different patients.

As in my thesis, we are working with second case, where we are looking for molecular insights into the progression of severe influenza A virus disease, the data are given as absolute intensities with respect to a common reference sample (i.e., mock infected). To identify DEGs, with respect to certain biological question, one can perform a suitable statistical test for each probes (genes). The choice of test statistics depends on the nature of the experimental data and what biological question we are asking for. In R-package *limma*, the basic statistics used for significance analysis is the moderated

$t-$statistics, which is computed for each probe and for each contrast. The R-package *limma*, provides *topTable*() as well as *decideTest*() functions, which summarises the results of the linear model, perform hypothesis tests and adjust the $p-$value.

We used *decideTest*() function to estimates the differentially expressed genes through out the whole infection process. In total, 1628 DEGs were identified during the whole infection process (see Fig. 3.4), while the list of genes is given in the appendix table B.1.

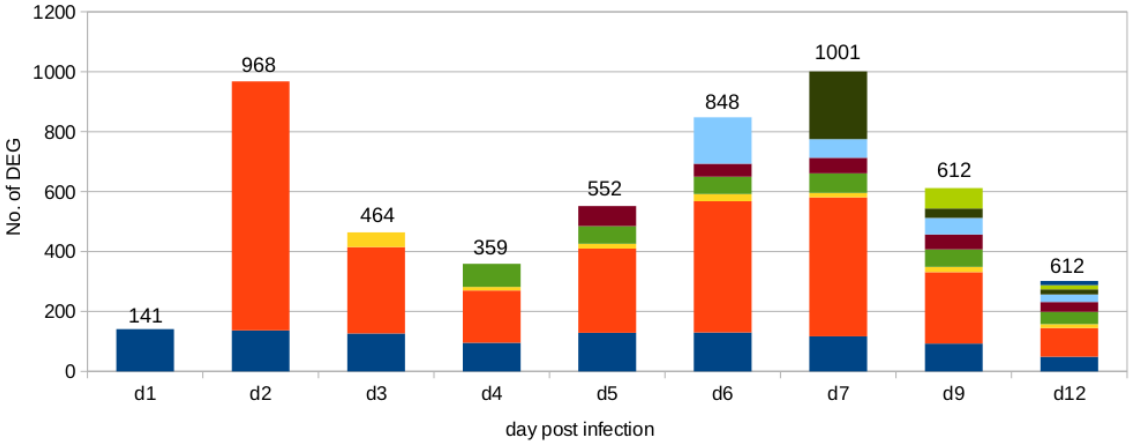


FIGURE 3.4: Gene expression changes in the lungs of mice infected with influenza A virus mp Jena/5258. A total of 1628 differentially expressed genes were identified during the infection process of 12 days compared to controls.The new colour at each days represent unique DEGs compared to the previous days while the same old colour represent that the presence of DEGs from the previous days, for example, dark blue represents the DEGs presents at day 1 and continue to presents through the infection process with different frequency while the orange colour at day 2p.i., represents new unique DEGs present at day 2 only but not before.

Most of them, i.e. 968 genes are differentially expressed (both up- and down-regulated) with respect to control data already at day 2 p.i. indicating a strong innate immune response. Genes which are strongly expressed during this early phase of infection indicate the induction of cellular factors involved in the detection of virus-associated molecular pattern (Appendix table B.2). Some of them are known to trigger the activation of innate immune response, proteins involved in antiviral response (e.g. $Rsad2$, $Mx2$, $Irf7$, $TLRs$) and chemokine/cytokine expression (e.g. $Ccl7$, $Cxcl1$, $Cxcl10$) [91].

During the following two days p.i. the number of DEGs decreased markedly with less than 400 DEGs, day 4 p.i., compared with the control. Then, a further strongly increased transcriptional response was observed with 1001 DEGs at day 7 p.i.,. Genes which are strongly expressed during this phase represent chemokine/cytokine response (e.g. $Ccl7$, $Cxcl-9/-10$). Finally, during recovery a decrease of the number of DEGs to 308 at day 12 p.i. was observed. The detailed information of genes having differential

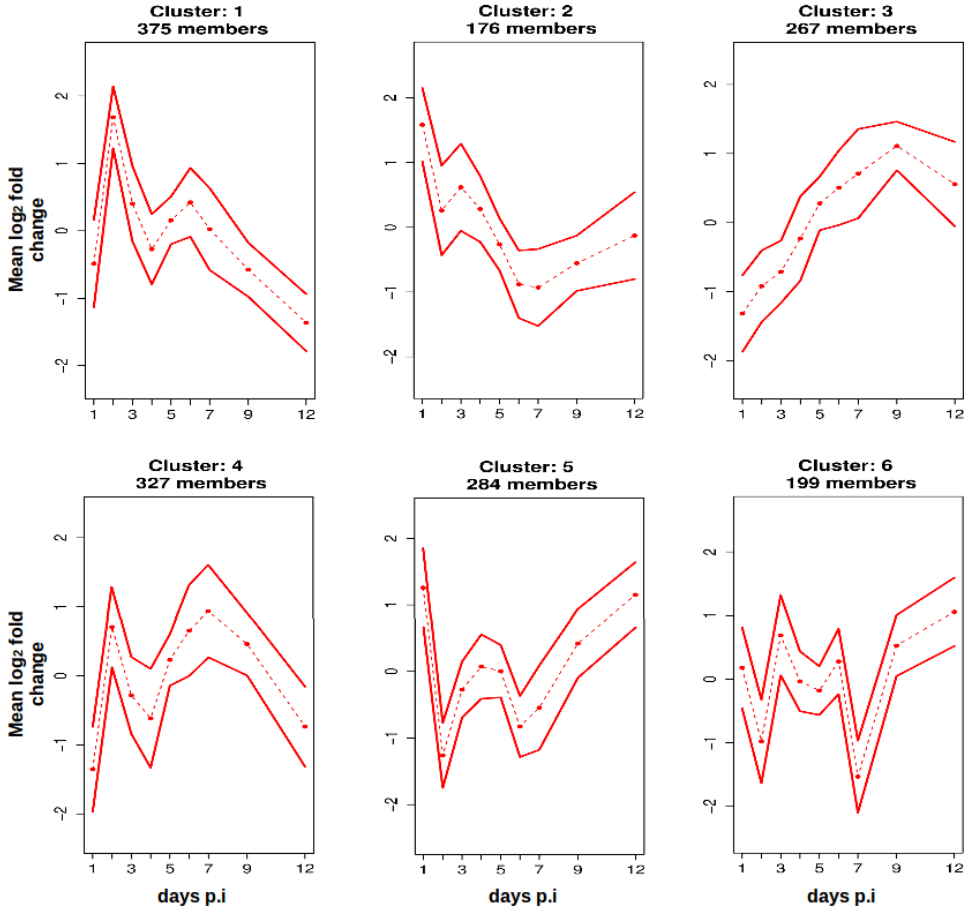
expression in comparison with control and for each day is provided in the Appendix table B.1.

In order to identify candidate genes of special interest for inference of a focused small-scale network, significantly overrepresented functional categories for 1628 DEGs were identified using the DAVID tool [92, 93] (Appendix table B.2). A number of 20 DEGs belong to the overrepresented GO term "response to the virus", with p-value of 1×10^{-5}, were considered. Four of them have pairwise very similar functions: $Oas1a/Oas1b$, $Mx1/Mx2$, $Ddx58/Ifih1$, and $H2Q7/H2Q8$. Thus, 16 genes belonging to the GO term "response to the virus" were included in the set of genes selected for regulatory network construction (Table 3.4).

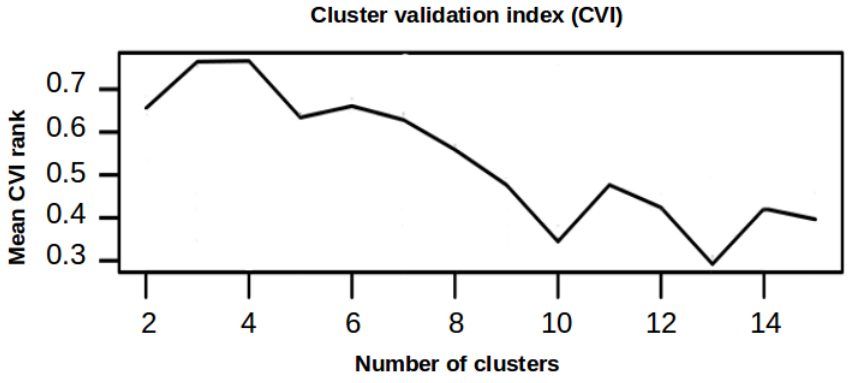
3.2.3 Cluster Analysis

In order to group genes with similar expression patterns over the time course of infection and to check in which extent the expression profiles of selected 20 genes are representative for all DEGs, we performed fuzzy c-means clustering [83] using all DEGs (see Methods 2.4.4.2). The number of clusters was optimised as previously described [84]. We found that six clusters optimally represent the dynamics in the dataset (Fig.'s 3.5a and 3.5b). Four of the 6 clusters, i.e. the clusters 1, 4, 5 and 6 comprising 1,185 DEGs (73% of all DEGs), are characterized by a biphasic temporal expression pattern over the time course of infection showing strong up- and down-regulation in the early phase which is followed by a stagnant recovery phase in the middle and further strong up- and down-regulation in the later phase of infection, which is ultimately followed by final recovery. Two of the four biphasic clusters, i.e. clusters 1 and 4, are being associated with the overrepresented GO category 'Immune response' (See appendix table B.2)

Cluster 1 contains 375 genes that were strongly activated in the early phase of the host response with a strong first peak at day 2 p.i. and a second peak of lower intense at day 6 p.i. and then down regulation from day 7 p.i. on. Interesting genes involved in this cluster are $Ifnb1$, $Il1a$, $Il1b$, $Myd88$, $Rsad2$, $Oas1a$, $Mx1$, which represent the detection of virus and activation of innate immune response by type I IFNs. The GO enrichment analysis of genes belonging to cluster 1 shows that this cluster mainly comprised the GO terms 'inflammatory response' and "response to virus", which represents the early induction of innate immune response in the host system. **Cluster 4** is associated with 327 genes that were strongly activated in the later phase of the host response with a second strong peak at day 7 p.i. and then the down regulation from day 9 p.i. on. Genes within this cluster are e.g. $Ccl-3/-4/-5$, $Cxcl-9/-13/-16$, $Stat-1/-4$, $Socs1$, $Nkfbid$.



(A)



(B)

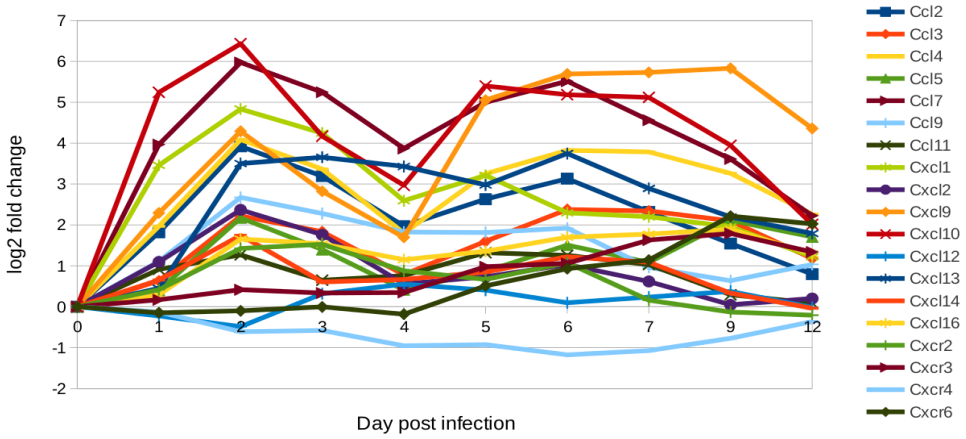
FIGURE 3.5: Fuzzy c-means clustering of DEGs reveals 6 cluster, (A) Mean expression profile with standard deviation of 6 clusters, identified by fuzzy c-means clustering. The x-axis represents the days post infection whereas the y-axis represents the mean scaled log2 fold change expression for the cluster., and (B) shows shows the mean cluster validation index (CVI) values based on different cluster validation criteria, for example Dunn index and Davis–Bouldin index (for details see [25]). The x-axis represents the number of cluster estimated and y-axis represents the corresponding CVI values. Higher value of CVI represents good estimates of the cluster number. We found 6 numbers of clusters quite stable with respect to the CVI value and they also covers most of the expression pattern of the DEGs over infection process.

They are assigned to the overrepresented GO terms 'Immune response' , 'positive regulation of immune response', inflammatory response', 'response to wounding' or 'regulation of cytokine production'. The cytokine and chemokine production along with their regulation by *Stat* and $Nkfb$ signalling may lead to higher infiltration of immune and inflammatory cells to the lung which may leads to acute lung injury which would be the basis of severe pathology. Severe complications arising from highly pathogenic influenza viruses are often associated with rapid and massive inflammatory cell infiltration. The complex interplay between cytokines and chemokines is a key step in shaping the host's immune response against invading pathogens including influenza viruses. In particular, the virus induced expression of cytokines and chemokines controls the recruitment and activation of neutrophils, macrophages, eosinophils and further immune cells to sites of infection. Both cluster 1 and cluster 4 are enriched with many cytokines and chemokines genes (see Fig. 3.6a) with bipahsic behaviour, thereby posing a significant role for an excessive host response in severe influenza infection. These data corresponds with the mean lung histopathological score observed over time (see Fig. 3.7) [2].

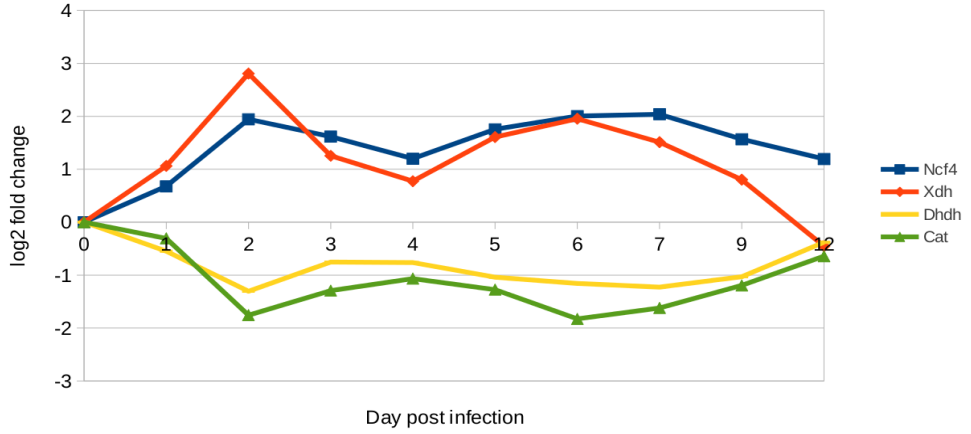
Cluster 5 and **cluster 6** comprise of 284 and 199 genes, respectively. Both clusters show two strong down-regulation peaks at days 2 and 7 p.i., which allocation coincides with temporal location of the peaks of the clusters 1 and 4. For cluster 5 the early peak is stronger than the second one, whereas the opposite for cluster 6. These two clusters are enriched with the GO terms 'cell projection morphogenesis', and 'cell morphogenesis' , 'epithelium development' respectively (Appendix table B.2). Biphasic expression profiles of genes belonging to the clusters 1 and 4 as well as 5 and 6 indicate also a strong expression pattern of ROS related genes. Higher expression of pro-oxidation genes such as $Ncf4$ and Xdh, and down-regulation of anti-oxidation genes such as $Dhdh$ and Cat has been found (see Fig. 3.6b). This strong pro-oxidation stimulation may lead to acute lung injury or pulmonary damage which further increase the severity of the disease.

Cluster 3 containing 267 genes shows up-regulation until a peak at day 9 p.i.. This cluster is enriched with the GO terms 'cell cycle and division', 'positive regulation of T-cell, 'lymphocytes and leukocytes activation' and 'positive regulation of immune system process', which shows strong adaptive immune response and stronger infiltration of immune cells (See appendix table B.2).

On the contrary to the pattern of cluster 3, **Cluster 2** consist of 176 genes showing down-regulation until day 7 p.i. GO enrichment analysis reveals that genes belonging to this cluster are associated with haemopoietic or lymphoid organ and blood vessel development, which clearly shows that with influenza infection the developmental processes goes down. The detailed GO enrichment analysis of all the 6 cluster is presented in appendix table B.2.



(A)



(B)

FIGURE 3.6: This figure illustrates expression profile of (A) various chemokines and their receptors, and (B) shows higher expression of pro-oxidation genes (Ncf4 and Xdh) and lower expression of anti-oxidation genes (DhDh,Cat). The y-axis covers the expression changes, represented as compared to the mock-infected control mice on a log2 scale, whereas the x-axis represents the days post infection.

3.2.4 Gene Regulatory Network Construction

3.2.4.1 Identification of Candidate Genes

As the number of possible networks structure increases exponentially with the number of genes, a small number of DEGs was selected to be included in the network reconstruction algorithm. In order to identify candidate genes of special interest for inference of a focused small-scale network, significantly overrepresented functional categories for 1628 DEGs were identified using the DAVID tool [92, 93]. A number of 20 DEGs belonging to the overrepresented GO term 'response to the virus', with p-value of 1×10^{-5}, were considered (See appendix table B.3). Four of them have pairwise very similar functions: $Oas1a/Oas1b$, $Mx1/Mx2$, $Ddx58/Ifih1$, and $H2Q7/H2Q8$. Thus, the majority of

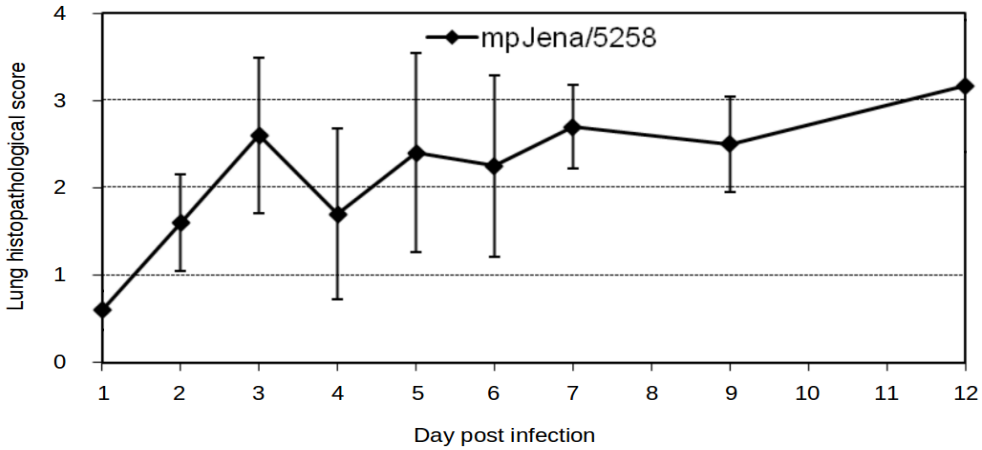


FIGURE 3.7: It shows the histopathological score of lung of mice infected with Influenza A virus, mp Jena/5258, as already published by Seidel et al. [2]. The x-axis shows the progression of infection process over days, while y-axis represents the histopathological score. Black square represents the averaged value of the histopathological score while the bars represents the standard deviation across different mice.

these 20 genes include the 16 genes assigned to the enriched GO category 'response to virus'. In addition, we included four genes ($Stat-1/-3$, $Irf1$ and $Socs1$), which were shown to play a crucial role in feedforward and feedback inhibition of interferon and TLR signalling during macrophage activation [94] (Table 3.4).

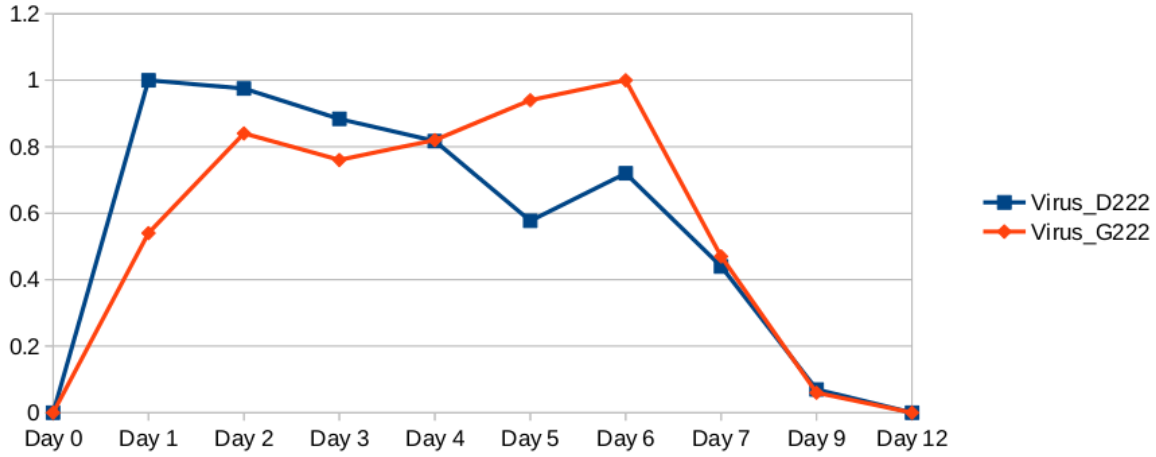


FIGURE 3.8: The perturbation factor values were calculated from the already published data of Seidel et al [2]. The values are calculated by the viral titre values given by Seidel [2], multiplied by the percentage value of quasispecies mp Jena/5258 (D/G) virus. The x-axis represents the infection process over days,(days post infection), while the y-axis represents the scaled viral titre value of mpJena/5258 D/G quasispecies. The blue solid line represents mpJena/5258 D222 variant while Orange solid line represents mpJena/5258 G222 variant.

TABLE 3.4: The 20 DEGs selected for network inference, along with their association with different clusters.

Genes	Function	GO Category other than "response to virus"	Cluster Association
Ddx58 : Dead box protein (Asp-Glu-Ala-Asp)	1) Plays a major role in sensing viral infection and activation of antiviral responses including type-1 IFN and pro-inflammatory cytokines. 2) It is involved in viral double-stranded (ds) RNA recognition and the regulation of immune response.	Innate immune response. +ve regulation of Ifn-alpha/-beta production. response to exogenous dsRNA. RIG-I signalling pathway.	1
Eif2ak2: Eukaryotic Translation Initiation Factor 2-Alpha Kinase	1) serine/threonine protein kinase that is activated by autophosphorylation after binding to ds-RNA. 2) Also involved in the regulation of signal transduction, apoptosis, cell proliferation and differentiation. 3) Regulates various signalling pathways like NF-Kappa-b, insulin-signalling pathway and importantly, can regulates various Tfs like JUN, Stat1, Stat3, Irf1, Atf3.	-ve regulation of viral genome replication. +ve regulation of chemokine n cytokine regulation. +ve regulation of NF-kappa-b signalling.	1
H2-Q7 or HLA-A: MHC Class I Antigen A	1) Involved in the presentation of foreign antigens to the immune system via MHC class I.	Antigen processing and presentation**.** +ve regulation of T cell mediated cytotoxicity.	4
Ifi27l2a: Ifi27L2 IFN, alpha inducible protein 27-like 2	1) Important paralog of this gene is IfI27, which mediates IFN-induced apoptosis.	Aging.	1
Ifnb1: IFN beta 1 Type 1 IFN	1) Mainly involved in the signalling of Toll-like receptor signalling pathway and Immune response IFN alpha/beta signalling pathway.	Adaptive immune response. Cytokine activity. Defense response to virus. B-cell proliferation and differentiation.	1
Ifng: IFN gamma Type 2 IFN	1) The protein encoded is a soluble cytokine with antiviral, immunoregulatory and anti-tumor properties and is a potent activator of macrophages.	Cytokine activity. CD8-+ve, alpha-beta T-cell differentiation. +ve regulation of IL-6, IL-12 biosynthetic process and T-cell proliferation.	3
Irf1: IFN regulatory factor 1	1) Activator of IFN alpha and beta transcription and also transcription activators of IFN induced genes like stat1.	+ve regulation of type 1 IFN & IL-12. IFN-g mediated signalling	4

	2) Plays an important role in immune response directly affecting NK maturation and activity, macrophage production of IL12, Th1 development and maturation of CD8+ T-cells	regulation of Myd88-dependent TLR signalling. CD8-+ve, alpha-beta T-cell differentiation.	
Irf-7	1) play a role in the transcriptional activation of virus-inducible cellular genes, including Ifnb. 2) Can efficiently activate both the IFN-beta and the IFN-alpha genes and mediate their induction via both the virus-activated, MyD88-independent pathway and the TLR-activated, MyD88-dependent pathway.	+ve regulation of IFN- alpha/-beta production. Regulation of Myd88 dependent TLR signalling pathway.	1
Lcn2: Lipocalin 2	1) Iron trafficking proteins involved in multiple processes such as apoptosis (due to IL3 deprivation), innate immunity and renal development.	Extrinsic apoptotic signalling pathway. Iron ion binding.	4
Mx1: Myxovirus (Influenza virus) resistance 1 Disease associated: Influenza virus	1) GTP metabolising proteins that participate in cellular antiviral response. 2) Inhibits IAV replication by decreasing or delaying NP synthesis and by blocking endocytic traffic of incoming virus particles.	Innate immune response. Response to virus.	1
Myd88: Myeloid differentiation primary response 88	1) Adapter protein involved in the Toll-like receptor and IL-1 receptor signalling pathway in the innate immune response. 2) Acts via IRAK1, IRAK2, IRF7 and TRAF6, leading to NF-kappa-B activation, cytokine secretion and the inflammatory response.	+ve regulation of IL-6, IL-17, IL-23. +ve regulation of NF-kappa-B, JNK signalling. Regulation of inflammatory response.	1
Oas1a n: Oas1: 2'-5' Oligoadenylate Synthetase 1 P52 Isoform	1) Interferon-induced, dsRNA - activated antiviral enzyme which plays a critical role in cellular innate antiviral response. 2) Mutations in this gene have been associated with host susceptibility to viral infection.	-ve regulation of viral process. 2'-5'- oligoadenylate synthetase activity.	1
Ptprc : Protein Tyrosine Phosphatase Receptor	1) PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitosis, and oncogenic transformation. 2) Acts as a positive regulator of T-cell coactivation upon binding to DPP4	Activation of MAPK activity. -ve regulation of viral genome replication. -ve regulation of protein kinase activity.	3
Rsad2: Radical S-	1) a major role in the cell antiviral state	CD4+, alpha-beta T-cell	1

Adenosyl Methionine Domain-Containing Protein 2	induced by type I and type II interferon. 2) Displays antiviral activity against influenza A virus by inhibiting the budding of the virus from the plasma membrane by disturbing the lipid rafts.	activation and proliferation. +ve regulation of TLR-7 and TLR-9 signalling pathway.	
Socs1: Suppressor of Cytokine Signaling 1	1) SOCS family proteins form part of a classical negative feedback system that regulates cytokine signal transduction. 2) SOCS1 is involved in negative regulation of cytokines that signal through the JAK/STAT3 pathway.	Kinase inhibitor activity. JAK-STAT cascade. -ve regulation of insulin-receptor signalling pathway	4
Stat1: Signal Transducer And Activator of Transcription 1	1) Involved in positive feed forward loop of IFN-gamma activation. 2) Acts as transcription activators and mediates the expression of a variety of genes important for cell viability in response to different cell stimuli and pathogens.	Cytokine mediated signalling pathway. IFN-gamma mediated signalling pathway Type-1 IFN signalling pathway. Regulation of NF-KAPPA-B, JAK-STAT cascade	4
Stat3: Signal Transducer And Activator of Transcription 3	1) Involved in negative feedback loop for IFN-gamma in combination with Socs1 2) plays a key role in many cellular processes like cell growth and apoptosis.	Involved in JAK-STAT cascade. -ve regulation of cell death & cell proliferation	1
Tlr3: Toll like receptor- 3	1) They recognize pathogen associated molecular patterns (PAMPs) and mediates the production of various cytokines. 2) It recognize dsRNA associated with viral infection and induces the activation of NF-kappa-B and production of type I interferons.	+ve regulation of NF-KAPPA-B, IFN-beta, IL-6, IL-12, JNK-cascade. +ve regulation of type-I and type-III IFNs. +ve regulation of NF-kappa-B signalling.	1
Tlr7: Toll like receptor- 7	1) Plays a role in pathogen recognition and activation of innate immunity. 2) Predominantly expressed in lung, placenta and spleen.	+ve regulation of IL-6, IFN-alpha Defense response to virus. +ve regulation of chemokine & IL-8 production and IFN-alpha/- beta synthesis	1
Tgtp1: T cell specific GTPase 1	1) It performs GTPase activity.	Response to IFN- alpha/-gamma. cellular response to IFN- beta. GTP catabolic process.	4

For network modelling we used two input perturbation functions defined by the two different quasispecies of mpJena/5258 calculated from the data shown by Nora [2] and shown in Fig. 3.8.

3.2.4.2 Extraction and Integration of Prior Knowledge: Pathway Studio 9.0

Prior knowledge comprising known and possible relevant interactions was extracted from various literature and other database sources, For example Pathway studio [95]. The prior knowledge used for network reconstruction from the pathway studio includes only of type direct interaction as well as expression. Several studies have shown that the inclusion of prior knowledge improves the reliability of the network inference approach [85–87, 96–99]. A confidence score was assigned to each prior knowledge used for the network inference based on the source of the data. High score 0.5 was given for higher number ($\geq$ 70) of supporting literature references, low score 0.1 for low number (between 5 and 10) of supporting literature references and median score 0.25 for numbers of references between 10 and 70 (See table 3.5). The prior knowledge was softly integrated during the network inference i.e. flexibly. Since different data sources might be contradictory, it is advantageous to softly integrate them during the modelling process.

3.2.4.3 Gene Regulatory Network of Murine Influenza Infection

In my thesis, I developed a system of ordinary differential equations (ODE) for constructing a gene regulatory network by integrating gene expression profile along with prior knowledge, extracted from Pathway Studio 9.0 [95]. Table 3.5 shows the prior knowledge information, and their corresponding confidence score used for network inference. As the complexity of network structure increases exponentially with the number of genes involved, a small number of 20 DEGs were selected to be included in the network reconstruction. The majority of these 20 genes include the 16 genes assigned to the enriched GO category "response to virus" (see Table 3.4). In addition, we included four genes ($Stat-1/-3$, $Irf1$ and $Socs1$), which were shown to play a crucial role in feedforward and feedback inhibition of interferon and TLR signalling during macrophage activation [94]. Fig. 3.9 shows the flowchart for data analysis, gene selection and the construction of regulatory network from the expression profiles of the selected genes. The final stable regulatory network is shown in Fig. 3.10, which simulated kinetics fit well to the expression profiles of the selected DEGs (See appendix Fig. A.1).

Overall, the network consists of two influenza variants that define the two input (perturbation) functions as shown in Fig. 3.8 and 37 edges, representing 27 gene-to-gene

Table 3.5: Prior Knowledge with the corresponding confidence score

	Relation		Type	# of References	Confidence Score
	Node1	Node2			
1	$Ifng$	$H2-Q7$	+ve	100	0.5
2	$Ifng$	$Irf1$	+ve	100	0.5
3	$Ifng$	$Stat1$	+ve	100	0.5
4	$Tlr3$	$Ifnb1$	+ve	100	0.5
5	$Ddx58$	$Ifnb1$	+ve	74	0.5
6	$Ifng$	$Socs1$	+ve	72	0.5
7	$Eif2ak2$	$Ifnb1$	+ve	11	0.25
8	$Ifnb1$	$Eif2ak2$	+ve	11	0.25
9	$Ifnb1$	$Mx1$	+ve	20	0.25
10	$Ifnb1$	$Oas1$	+ve	11	0.25
11	$Ifnb1$	$Stat1$	+ve	19	0.25
12	$Ifng$	$Myd88$	+ve	14	0.25
13	$Ifng$	$Oas1$	+ve	13	0.25
14	$Ifng$	$Tlr3$	+ve	13	0.25
15	$Myd88$	$Ifng$	+ve	19	0.25
16	$Myd88$	$Irf7$	+ve	30	0.25
17	$Ptprc$	$Ifng$	+ve	12	0.25
18	$Tlr3$	$Ifng$	+ve	20	0.25
19	$Tlr3$	$Myd88$	+ve	21	0.25
20	$Tlr7$	$Ifnb1$	+ve	15	0.25
21	$Tlr7$	$Ifng$	+ve	15	0.25
22	$Tlr7$	$Myd88$	+ve	23	0.25
23	$Tlr7$	$Tlr3$	+ve	13	0.25
24	$Eif2ak2$	$Ifng$	-ve	7	0.10
25	$Eif2ak2$	$Irf1$	+ve	9	0.10
26	$Ifnb1$	$H2-Q7$	+ve	7	0.10
27	$Ifnb1$	$Rsad2$	+ve	5	0.10
28	$Ifnb1$	$Socs1$	+ve	9	0.10
29	$Ifng$	$Ddx58$	+ve	10	0.10
30	$Ifng$	$Eif2ak2$	+ve	9	0.10
31	$Ifng$	$Irf7$	+ve	6	0.10
32	$Ifng$	$Tgtp1$	+ve	7	0.10
33	$Irf1$	$Eif2ak2$	+ve	6	0.10
34	$Irf1$	$Tlr3$	+ve	6	0.10
35	$Myd88$	$Irf1$	+ve	7	0.10

interactions and 10 influences of the influenza variants on the gene expression. A subset of 15 of the gene-to-gene interactions (coloured green in Fig. 3.10) were previously reported and are introduced as prior knowledge as shown in Table 3.5 extracted by Pathway Studio [95]. The remaining 12 gene-to-gene interactions were either novel interactions predicted by the inferred network or already known but not included in the prior knowledge. Interestingly, we found among these 12 predicted edges two gene-to-gene interactions which were previously described in literature, but not used as in prior knowledge for the network inference: $Ddx58$ positively activating $Stat3$ and $Mx1$

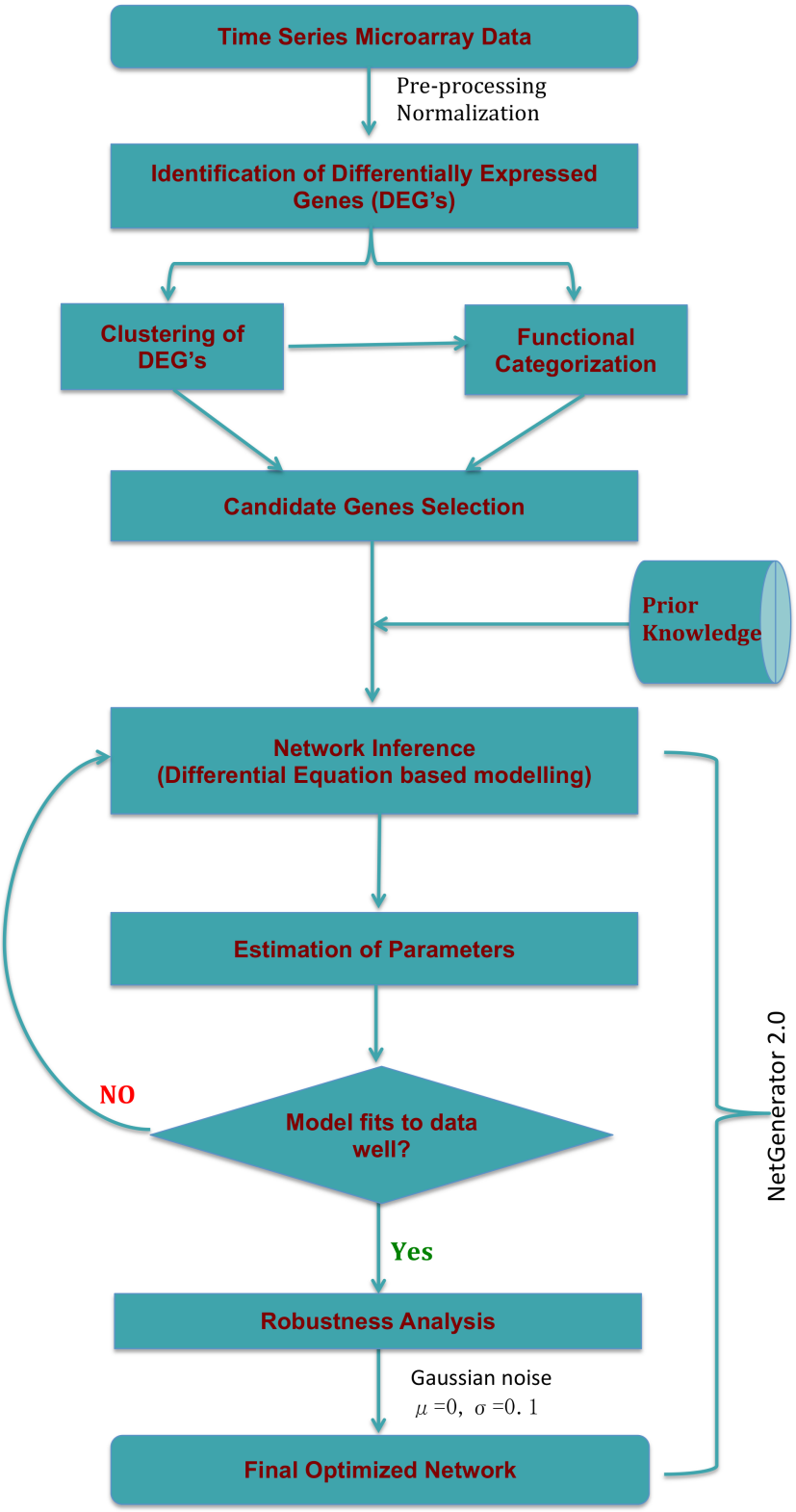


FIGURE 3.9: Flowchart showing the analysis of microarray time series data utilized for the construction of gene regulatory networks using the NetGenerator tool [87].

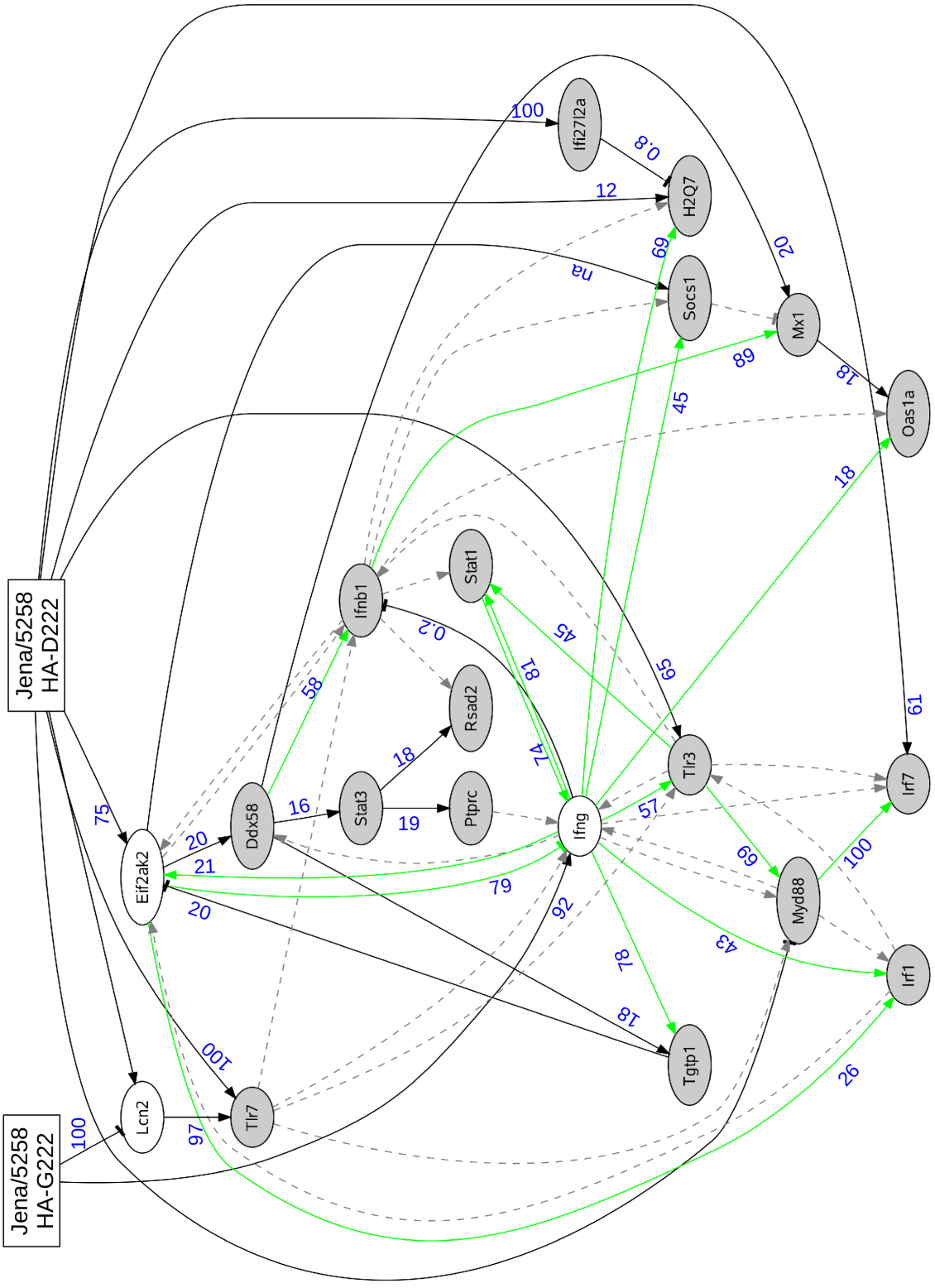
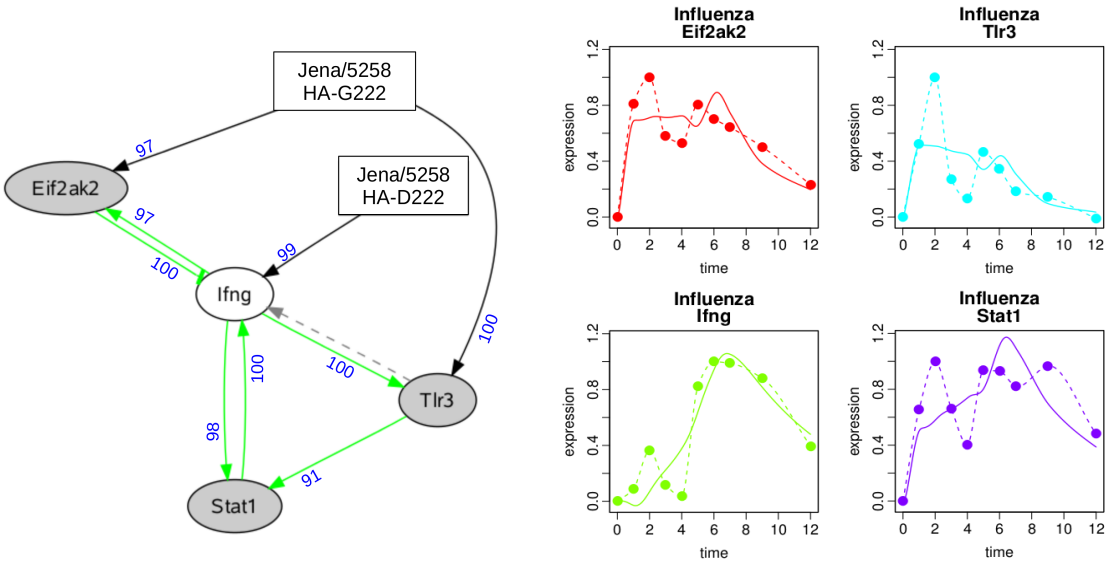


FIGURE 3.10: Gene-regulatory network prediction from high throughput time series microarray data for 20 DEGs. Where "mpJena/5258 D222" and "mpJena/5258 G222" represents the two influenza variants found by Seidel et al. [2]. Black edges represents the newly predicted ones, green edges represent edges supported by the prior knowledge and confirmed by the expression data-based network inference and grey dotted edges represent prior knowledge not included in the network prediction based on the measured gene expression profiles. The number represent the percentage of robustness after 500 iteration. Arrow-head represents activation or positive regulation while bar-head represents repression or negative regulation (that may also represent indirect interaction).

[100–104]. This result of validation by external data from literature confirms the predictive power of modelling. Among the 37 edges, we found 5 repressions or inhibitory effects and 32 activations among the overall network.

3.2.4.4 Inflammatory Sub-network is Robust

As a part of the inferred network shown in Fig. 3.10, there is one positive feedback loop formed by the interaction of the genes $Tlr3 - Ifng - Stat1$. To better understand the regulatory mechanisms of hyper-responsive behaviour induced by severe A(H1N1)pdm09 infection, the robustness of this sub-network involving the positive feedback loop was investigated. We constructed two differential sub-networks involving four genes ($Tlr3-$



(A) (B)

FIGURE 3.11: Network inference and simulated kinetics for 4 genes, (A) Gene-regulatory network prediction from high throughput time series microarray data by NetGenerator 2.0 involving 4 genes which were connected to the positive feedback loop. Where "mpJena/5258 D222" and "mpJena/5258 G222" represents the two influenza variants found by Seidel et al. [2]. Black edges represents the newly predicted ones, green edges represents edges supported by the prior knowledge and confirmed by the expression data-based network inference and grey dotted edges represent prior knowledge not included in the network prediction based on the measured gene expression profiles. The number represent the percentage of robustness after 500 iteration. Arrow-head represents activation or positive regulation while bar-head represents repression or negative regulation (that may also represent indirect interaction). and (B) showing how well the data was fitted to the observed gene expression data of 4 DEGs ($Ifng$, $Tlr3$, $Stat1$, and $Eif2ak2$) that are connected with the positive feedback loop as shown in result section. Dotted line shows the observed gene expression data for a gene over a period of 12 days, while the solid line shows the model fit. The x-axis shows the progression of infection over days, (day post infection), while the y-axis represent the gene expression changes (logarithmised and scaled).

$Ifng-Stat1-Eif2ak2$) and 6 genes ($Tlr3-Ifng-Stat1-Eif2ak2-Socs1-Ifi27l2a$) adding one or three genes, respectively, that are connected via robust edges with the 3-gene-loop.

For the robustness analysis, we used the same algorithm as described above to construct the network from the gene expression data (See Methods 2.4.4.3). The robustness analysis for the 4-gene-sub-network was performed by adding Gaussian noise of mean 0 and standard deviation (sd) of 0.01 and repeated network inference procedure for 500 times. The robustness analysis reveals that almost all the interactions were robust, i.e. the edges were recovered with a rate of 90%. The final network and simulation is shown in Fig. 3.11. The simulated profiles show good agreement with the measured gene expression data, which indicates the reliability and predictability of 4-gene-model shown in Fig. 3.11b. Similar analysis was done for the 6-gene-subnetwork with sd =0.01. The robustness analysis also reveals remarkably high percentage (recovery rate $\geq$90%) for almost all the interactions involved, with an exception of an edge representing the influence of Influenza-2 variant to the gene $Socs1$ (62%) The final 6-gene-subnetwork and simulation is shown in appendix Fig.'s A.2 and A.3 respectively. These results demonstrates that the positive feedback loop formed by the genes $Tlr3 - Ifng - Stat1$ is quite stable.

3.2.4.5 High Association Between Phenotypic and Genotypic Data

To find the association between the phenotype of the disease course with the corresponding whole genome expression data, we used multiple linear regression modelling. Here, the observed phenotypic data quantified by the clinical (symptom) score S from Manchanda et al. [22] was used as the dependent response variable. The scaled gene expression ratios values of the four genes of the 4-gene-subnetwork (i.e., $Ifng, Stat1, Tlr3, Eif2ak2$) were used as dependent (explanatory) variables. The linear regression model is shown in the following equation and was fit by lm function in R:

$$\mathbf{fit} \;=\; \mathbf{lm}(\mathbf{formula} = \mathbf{S} \approx \mathbf{Ifng} + \mathbf{Stat1} + \mathbf{Tlr3} + \mathbf{Eif2ak2} - \mathbf{1})$$

Interestingly, we found that the individual influence of the four individual independent variables have no significance (p-values ≥ 0.3) whereas a significant influence were found for the combined influence of the four genes on the clinical score (overall p-value 0.024). These results suggests that the phenotype of the disease is controlled by overall interactions between independent variables rather their individual expression.

Chapter 4

Discussion

4.1 Small-scale Mathematical Models for Influenza Infection

The main focus of this study was to quantify the kinetics of influenza caused by different viral strains based on a single model by only few most important parameters. With this aim, a dynamic model of influenza in mice was established and fitted to experimental data representing virus-induced clinical symptoms. Three non-fixed model parameters were proven to be identifiable and to be suitable to quantitatively characterize the four different virus strains under study: (i) the viral replication and infection rate (α), (ii) the maximum primary pathogenicity (k_p), and (iii) the rate of early activation of the immune system (γ). However, the model structure is hypothetic. Follow-up experiments, including molecular parameters and genome-wide transcriptome analysis are ongoing to discriminate between alternative models and to verify the hypothesis that the second peak of biphasic course of infection is caused mainly by inflammation. Consequently, the models suggested in the present study describe the course of influenza based on virus infection and on virus-induced immune response, including pro- as well as an anti-inflammatory response. Finally, they link the virus pathogenicity and/or inflammation with the observable clinical symptoms.

The small-scale model has some interesting features: First, it can describe the outcome of infection with four influenza virus strains. As a result, the model can be used to quantify the virulence of different strains by a set of eight parameters listed in tables 2.1 and 3.1. Three of them (α, k_p, γ) were found to be structurally as well as practically identifiable. The remaining five parameters are non-identifiable with the given experimental data. They were fixed, which no critical for four of them due to their low sensitivity (see Table 3.2). Only for the threshold parameter δ introduced by Kumar et al. [46], the practical

non-identifiability is an issue for further studies due to its high sensitivity for biphasic profiles of the clinical score.

The model can simulate and interpret the cause of different outcomes of disease after virus challenge in mice. At day 2 p.i., mice became ill independently from the virus strain and infection dose studied. Maximum clinical score was observed between days 3 and 6 after infection with Jena/5555, Bakum/1832 and Jena/2688; afterwards, mice recovered (i.e. monophasic disease kinetic). Furthermore, there is only low or no persistent inflammation. Obviously, the respective viruses have been eliminated efficiently by the immune response. In contrast to infection with Jena/5258 virus, a biphasic course of clinical symptoms was observed and simulated by the mathematical model. According to the established model, persistent inflammation contributes to the biphasic course of disease. Due to experimental design, it remains unclear whether the virus was eliminated or not at the second disease phase or peak. So there is still an open question, whether the second peak of the disease is solely due to the persistent inflammation or due to the existence of mutant influenza virus.

For modelling of the biphasic behaviour, a dynamic model introduced and theoretically studied by Kumar et al. [46], was used and reduced. To the best of our knowledge, in the present work, the fist time Kumar's model was fitted to experimental data by linking the strength of inflammation with clinical symptoms representing influenza infection in mice. In addition, the course of infection of A(H1N1)pdm09 virus as well as European swine H1N2 influenza virus (Bakum/1832) was modelled the first time.

Based on the model presented here, it can be assumed that the dynamics of influenza depends on both the virus pathogenicity as well as the adverse adaptive immune response that includes the pro- inflammatory cytokines

The model established in the present study can interpret and quantify the biphasic course of disease by a late inflammation phase. It allows quantifying the rate of pro- and anti-inflammatory response (ϵ and ρ respectively) as well as the conditions for the onset of inflammation (δ, ω) by a set of identified parameters. As suggested by Kumar et al. [46], it was hypothesised with equation 2.1 that the onset of inflammation (model variable I) is triggered by the immune defense (if the model variable D goes beyond the threshold value δ). The sensitivity of the pro-inflammatory response in dependence on the immune defense is quantified by the parameter $(1/\omega)$. These results warrant a more detailed biological characterization of pro-inflammatory response that is ongoing now. As published recently for a non-lethal infection of C57BL/6J mice with A/Puerto Rico/8/34 [59], a transcriptome analysis will help to clarify the role of factors of innate and adaptive host immune response in severe biphasic influenza in mice.

The observed clinical score was used for the presented model fit. These data are negatively correlated with the body weight (W; Fig. 2.4). Thus, the body weight could be used alternatively for model fit using a linear regression model as following: $(\mathbf{W = W_0 - a * P - b * I})$.

The model introduced by Canini and Carrat [48] was based on population analysis of viral kinetics and symptom's dynamics, which is similar to our study. The major limitation with their model was that data of immune response were not included. Their model was solely dependent on the pathogenicity of virus used. They assume that natural killer cells contribute to adverse inflammation, which supports the hypothesis of the small-scale model of the present study. In addition, the suggestion of Smith and Perelson [36] to exploit clinical score as easily attainable experimental data supports the approach used here.

The established model may help to quantify the dynamics of influenza induced by different virus strains and support the search for reasons for the severe, biphasic course of influenza in mice. Including antiviral treatment, drug administration could be optimized specifically for the virus strain used in further studies. The model might allow to direct antiviral studies, including the mathematical analysis of the use of antiviral compounds such as Oseltamivir for the treatment of pandemic influenza virus infections (as modeled e.g., for the HIV therapy by [105]).

4.2 Reconstruction of Gene Regulatory Network from Gene Expression Data.

Previous studies have shown that different isolates and variants of A(H1N1)pdm09 have quite different disease kinetics, some causing severe illness and others mild [1, 2]. In the present study, we infected BALB/c mice the once mouse-passaged A(H1N1)pdm09 isolate mpJena/5258 influenza virus causing severe, biphasic disease [2, 22]. Also, we analysed changes in the whole genome expression over a period of 12 days to (i) describe the dynamics of host response at the molecular level as well as gene regulatory level and (ii) to compare these patterns to the phenotypic level, in particular, to the virus-induced symptoms (clinical score). Our analysis represents, for the first time, a comprehensive analysis of the host response over the entire process of severe pandemic influenza A virus infection, including both innate and adaptive immune response. The results do not only reveal for the first-time biphasic gene expression profiles of host response against severe pIAV infection but also correlate with the phenotype of disease, i.e. clinical score as shown before [22]. In addition to genome expression analysis, we

used an ordinary differential equation based modelling approach to infer a gene regulatory network. To our knowledge, this is the first gene regulatory network involving two variants (HA-D222 and HA-G222 quasi-species) of A(H1N1)pdm09. The inferred network demonstrates a positive feedback loop between $Stat1, Ifng$, and $Tlr3$. This explains the hyper-responsive behaviour of $Ifng$ in presence of the HA-G222 variant, shown by the interaction of this variant directly with $Ifng$ in the regulatory network.

Recently, a high complexity for inflammatory networks, which is accompanied by high entropy and low free energy was shown, for a highly pathogenic influenza strain (H5N1) compared with the pandemic influenza strain (H1N1) of mild pathogenicity [23]. The complexity of the inflammatory network may likely contribute to the severity and lethality of disease associated with the highly pathogenic influenza strain. This is in accordance with our regulatory network prediction where we found a complex regulation of $Ifng$ (high indegree of $Ifng$, see Fig. 3.10). This is also shown by one study, where Ifng-deficient Tc1- or Tc2- CD8 effector cells lead to severe impairment in lung function with the elicitation of intense inflammatory response [106]. Furthermore, in another study of respiratory viral infection, it has been shown that the absence of $Ifng$ significantly reduced the lung pathology [107].

Another recent study [108], described an elevated activation of inflammatory signalling networks in lethal influenza infection compared with sub-lethal infection. In addition, a positive feedback chemokine-derived loop which regulates the pro-inflammatory response and also elevated levels of neutrophil's infiltration was shown. This observation is in concordance with our finding of a positive feedback loop between $Ifng$, $Stat1$ and $Tlr3$, where $Ifng$ represents the pro-inflammatory gene. As pro-inflammatory cytokines lack specific chemotactic activities, therefore, are the main factors enabling monocytes and T lymphocytes to migrate from the peripheral blood via the vascular endothelium into the site of inflammation/infection. Overall, our study provides strong evidence for the role of the positive feedback loop in triggering the pro-inflammatory response in the infected site to release immediately and to produce attractants that amplify the interstitial cellular inflammatory response.

A latest study [2] demonstrated that HA-222D/G quasi-species of A(H1N1)pdm09 facilitates intra host evolution accompanied with severe disease. According to our regulatory network prediction, the HA-G222 variant, which replicates with high viral titres in mouse trachea and lung tissues on day 6 p.i., [2], induces the strong $Ifng$ gene expression. This may also explains the hyper-responsive behaviour of $Ifng$. Possibly, this promotes the infiltration of inflammatory cells into the lung and thereby the observed lung histopathology.

A common feature of severe A(H1N1)pdm09 is the involvement of inflammatory immune response, characterized by activation of epithelial cells, macrophages and the recruitment and activation of neutrophils, eosinophils, monocytes and further immune cells. Inflammatory cells become activated and generate reactive oxygen species (ROS) [109]. ROS is highly reactive and leads to lipid peroxidation and increase of tissue permeability, which have been implicated in the pathogenesis of lung injury. The genome expression profile confirms a strong expression of pro-oxidation genes such as $Ncf4$ and Xdh along with, down regulation of anti-oxidation genes such as $Dhdh$ and Cat (Fig. 3.6b). The strong pro-oxidation stimulation may further increase lung injury and the severity of the disease. Furthermore, studies show that, reverse mutant viruses expressing HA and NA of the pandemic H1N1 influenza virus induced significantly higher levels of expression of $Ifng$, $IL1$, $IL6$, $Ccl4$, $Ccl5$, $Cxcl9$, and $Cxcl10$ in the lungs of infected mice [110, 111], supporting our study that the elevated pro-inflammatory response is an important hallmark of the host's response against pandemic influenza virus infection.

In addition, we also focused on the comparison between the virus-induced symptoms (clinical score and body weight changes [22]) and the gene expression profile. In fact, we have found for the first time a significant association of the virus-induced symptoms with whole genome.

In conclusion, the intra-host evolution of HA-222D/G quasi-species of mpJena/5258 elicits a complex response in infected mouse lung, characterized by a biphasic gene expression pattern. During the first phase, in particular, at day 2 p.i., when the HA-D222 variant dominated in the mouse lung and trachea [2], a strong innate immune response was detected with stronger expression of type I IFNs and pro-oxidation genes. In the second phase of disease, the evolution of the HA-222G variant was correlated with a hyper-responsive behaviour of $Ifng$ via the positive feedback loop $Tlr3 - Ifng - Stat1$. This finally leads to the stronger pro-inflammatory response demonstrated by the second peak in the expression profiles of the majority of differentially expressed genes at day 7 p.i. The expression profiles of the genes $Tlr3$, $Ifng$, $Stat1$, and $Eif2ak2$ were biphasic like the virus-induced symptoms, shown by Manchanda et al. [2, 22]. Additional, experimental validation of this predicted network will help to explain further its importance. The present study can serve as a significant basis for further exploring and comparing the host response between different pandemic influenza infections and developing control strategies.

Chapter 5

Summary

A model is a mathematical process of formulating generalised ideas of reality. The level of information depends upon the question asked and the scale at which underlying causative processes are studied. For example, in influenza infection process, most molecular events in host-pathogen interaction, e.g. types of immune cells involved, virus development inside the host, signalling pathways, are sometimes ignored because of the complexity of these processes. However, many of these processes are condensed into a single parameter in the antiviral immune defense. On the other hand, how the titres of the virus or any other pathogen change in an individual is studied, as that is what decides the diseased state of the host individual. Mathematical model of influenza dynamics helps to characterize, quantify and summarise the infection dynamics in the murine model. By clarifying rigorously the assumptions, the variables and the parameters, mathematical modelling allows better understanding of infection kinetics. These models further provide important conceptual results for e.g., replication rate, host defense response. For evident ethical and practical reasons, experiments in public health are often quite limited and impossible to perform. Thus, mathematical models appear to be as a cheap and efficient way to explore and test hypotheses.

In my thesis, I highlighted the importance of a small-scale mathematical model for characterization and quantification of different pandemic influenza A virus. By combining the genome wide gene expression profiling with the clinical symptom score to gain insights into molecular mechanism and characterization of key factors to severe influenza infection helps to develop therapeutic measures. The biological basis for the increased severity of some influenza viruses remains unclear. Unpredicted mutations which leads to intra-host evolution of quasi-species, and strong inflammation are important hallmarks of severe pandemic influenza infection. This emphasizes the need of development

of new mathematical model describing influenza kinetics of different strains, which can help in virus quantification based on their severity.

Apart from quantification of different influenza viruses, investigating the molecular mechanisms of severe pandemic IAV (pdmIAV) is of great importance in controlling the complications and reducing the pulmonary damage. In my thesis, I also performed a comprehensive analysis of whole genome expression changes of the lung of mice infected with the pH1N1 mpJena/5258 consisting HA-222D/G quasi-species to gain insights into pathogenesis at the transcriptional level. I analysed gene expression data from both, the innate and adaptive host immune response to identify key regulatory interactions during infection. A reverse engineering approach was applied to infer a gene regulatory network. Since the complexity of network structure increases with the number of genes, a small number of 20 differentially expressed genes (DEGs) were selected to be included in the network. The gene regulatory network of influenza infection shows a positive feedback loop. The mpJena/5258 influenza virus infection is associated with biphasic gene expression profile and a positive feedback mechanism of Ifng which correlates with the evolution of HA-222D/G quasi-species and leads to overwhelming immune response. Also, a significant correlation was also found between the co-expression action of three genes (Ifng, Stat1 and Tlr3) with severity of the disease.

In conclusion, a small-scale mathematical model of pandemic H1N1(2009) influenza A (A(H1N1)pdm09)) infection was developed with a special emphasis on the pro-inflammatory response. Simulation analysis of this model revealed that the pro - inflammatory response plays some adverse role on the disease condition, and it is specific to virus strain. For the first time, the dynamics of four different influenza A virus strains (three A(H1N1)pdm09 and one European swine H1N2) were modeled and quantified. The established model helps us in characterising different influenza virus strains into severe and mild categories based on few interpretable parameters. The model also shows that the pro-inflammatory response is very specific for the virus strains and thus, it plays a crucial role for optimisation of the therapeutic profile of drug administration that should be designed specifically for the virus strains. Based on the model quantification, we found that Jena/5258 was the most severe virus strain with having a biphasic clinical symptom score. Moreover, to investigate the molecular mechanism behind it, we analysed whole genome expression changes of the lung of mice infected with Jena/5258. The intra-host evolution of HA-222D/G quasi-species of mpJena/5258 elicits a complex response in infected mouse lung, characterized by a biphasic gene expression pattern. During the first phase, in particular, at day 2 p.i., when the HA-D222 variant dominated in the mouse lung, and tracheas [2], a strong innate immune response was detected with strong expression of type I IFNs and pro-oxidation genes. In the course of second phase

of disease, the evolution of the HA- 222G variant correlated with a hyper-responsive behaviour of Ifng via the positive feedback loop Tlr3 Ifng Stat1. This finally leads to the stronger pro-inflammatory response demonstrated by the second peak in the expression profiles of the majority of differentially expressed genes at day 7 p.i. The expression profiles of the genes Tlr3, Ifng, Stat1, and Eif2ak2 were biphasic like the virus-induced symptoms, shown by [2, 22]. This investigation can serve as a significant basis for further exploring and comparing the host response between pandemic influenza infection of different severity level and developing control strategies.

Appendix A

Appendix Figures

A.1 Simulated Kinetics for 20 Genes.

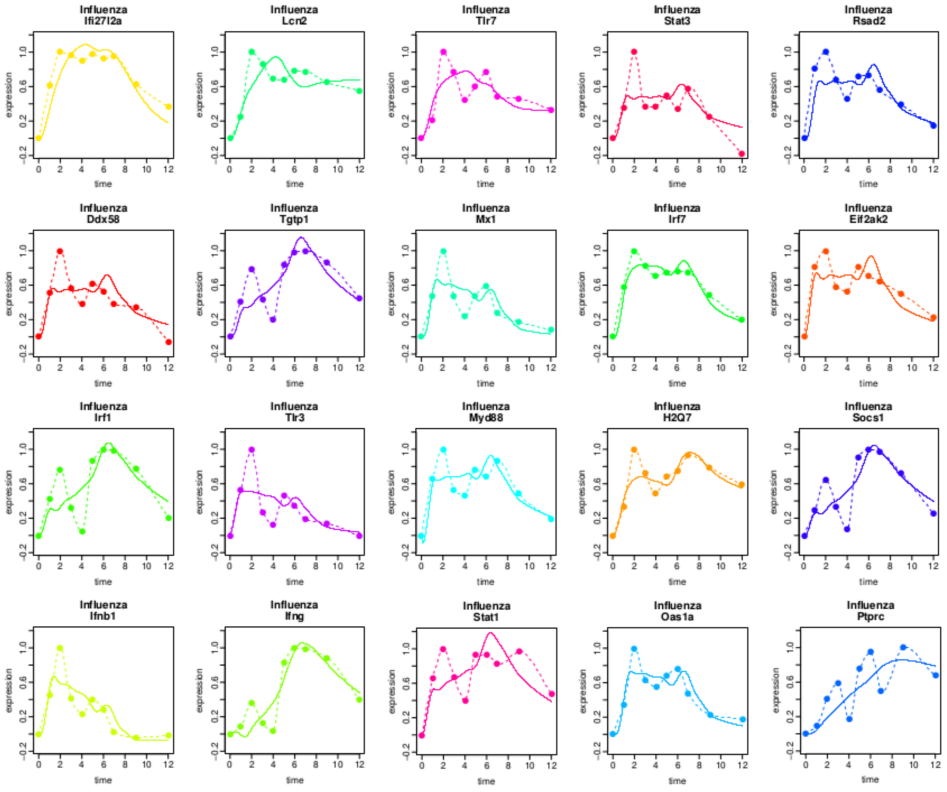


FIGURE A.1: It represents the output of the NetGenerator 2.0 R-package, showing how well the data was fitted to the observed gene expression data of 20 DEGs. Dotted line shows the observed gene expression data for a gene over a period of 12 days, while the solid line shows the model fit. The x-axis shows the progression of infection over days, (day post infection), while the y-axis represent the gene expression changes (logarithmised and scaled).

A.2 Gene Regulatory Network of 6 Genes.

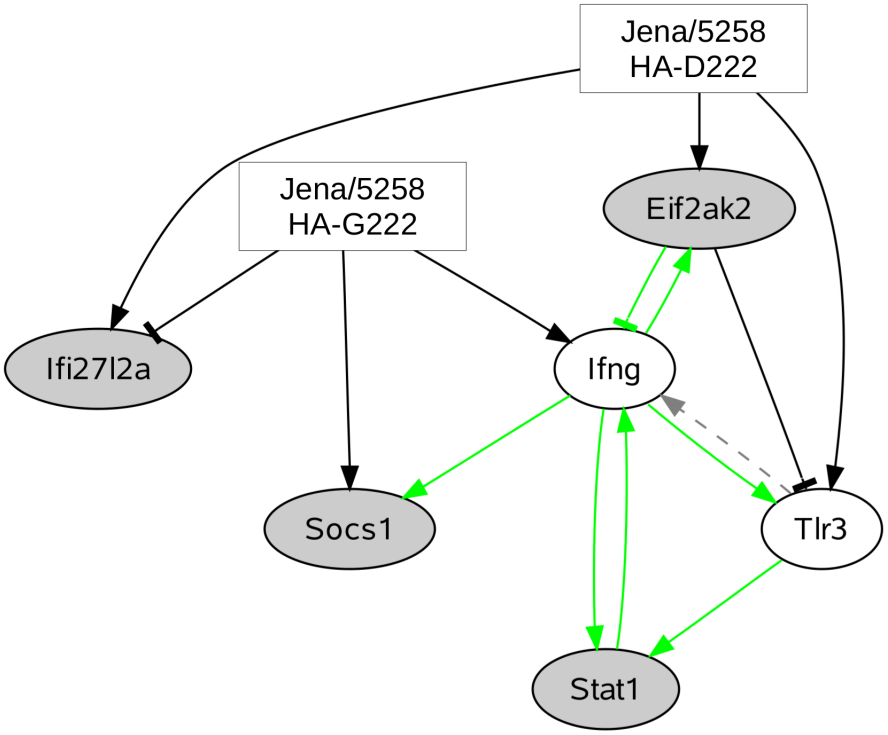


FIGURE A.2: Gene-regulatory network prediction from high throughput time series microarray data by NetGenerator 2.0 involving 6 genes which were connected to the positive feedback loop. "Jena/5258 HA-D222" and "Jena/5258 HA-G222" represents the two influenza variants found by Seidel et al. [2]. Black edges represents the newly predicted ones, green edges represents edges supported by the prior knowledge and confirmed by the expression data-based network inference and grey dotted edges represent prior knowledge not included in the network prediction based on the measured gene expression profiles. Arrow-head represents activation or positive regulation while bar-head represents repression or negative regulation (that may also represent indirect interaction).

A.3 Simulated Kinetics for 6 Genes.

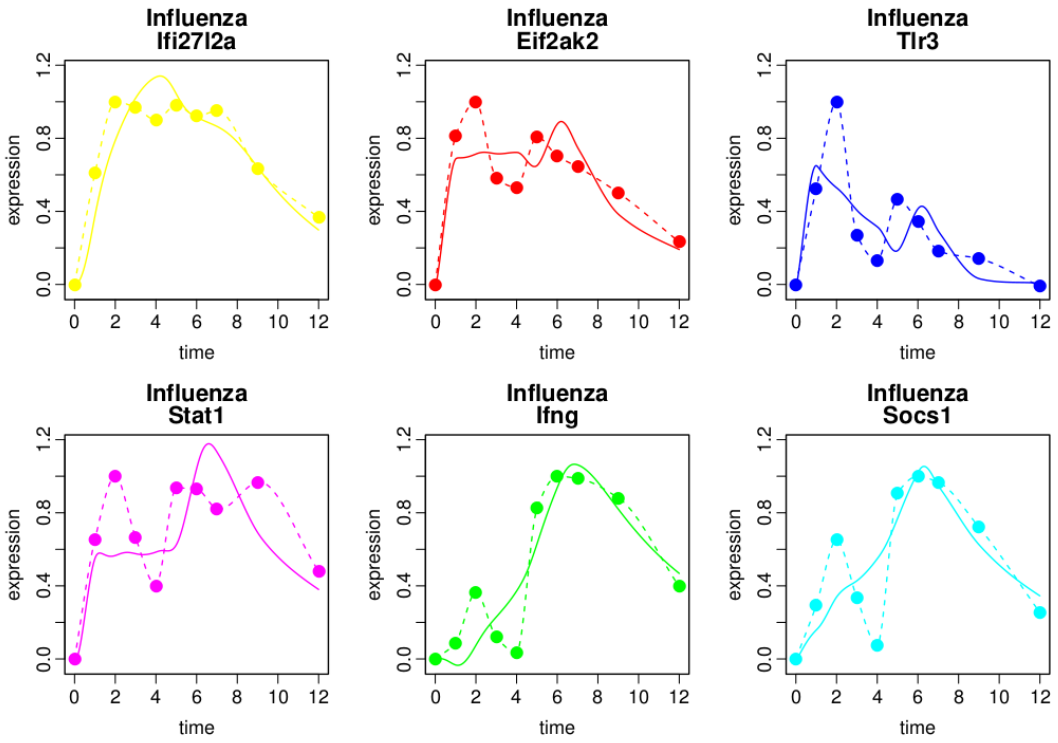


FIGURE A.3: Output of the NetGenerator 2.0 R-package, showing how well the data was fitted to the observed gene expression data of 6 DEGs ($Ifng$, $Tlr3$, $Stat1$, $Eif2ak2$, $Socs1$, and $Ifi27l2a$) that are connected with the feedback loop as shown in result section. Dotted line shows the observed gene expression data for a gene over a period of 12 days, while the solid line shows the model fit. The x-axis shows the progression of infection over days, (day post infection), while the y-axis represent the gene expression changes (logarithmised and scaled).

Appendix B

Appendix Tables

The attached CD contains the following data tables used in this thesis.

B.1 Differentially Expressed Genes over days.

The file **Table S1.csv** contains, gene expression changes in the lungs of mice infected with influenza A virus mp Jena/5258. A total of 1628 differentially expressed genes were identified during the infection process of 12 days compared to controls.

B.2 Gene Enrichment for 6 Clusters.

The file **Table S2.csv** contains, in total, six different cluster of DEGs with similar expression profile were identified by cluster analysis. The genes from individual clusters were mapped to significantly overrepresented functional categories were identified using DAVID tool.

B.3 Gene Enrichment for 1628 DEGs.

The file **Table S3.csv** contains, a total of 1628 differentially expressed genes were identified during the infection process of 12 days compared to controls, were mapped to significantly overrepresented functional categories were identified using DAVID tool.

Bibliography

[1] W. Zou, D. Chen, M. Xiong, J. Zhu, X. Lin, L. Wang, J. Zhang, L. Chen, H. Zhang, H. Chen, M. Chen, and M. Jin, "Insights into the increasing virulence of the swine-origin pandemic H1N1/2009 influenza virus," *Scientific reports*, vol. 3, p. 1601, Jan. 2013.

[2] N. Seidel, A. Sauerbrei, P. Wutzler, and M. Schmidtke, "Hemagglutinin 222D/G Polymorphism Facilitates Fast Intra-Host Evolution of Pandemic (H1N1) 2009 Influenza A Viruses," *PloS one*, vol. 9, p. e104233, Jan. 2014.

[3] World Health Organisation, "WHO —Influenza (Seasonal)," tech. rep., 2014.

[4] Robert Koch Institute, "Influenza associated mortalitity in Germany 1985-2010. Epidemiologisches Bulletin," tech. rep., Robert Koch Institute, Berlin, 2011.

[5] Robert Koch Institute, "Bericht zur Epidemiologie der Influenza in Deutschland Saison 2012/13," *Robert Koch-Institute, Berlin*, 2013.

[6] N. M. Bouvier and P. Palese, "The biology of influenza viruses," *Vaccine*, vol. 26 Suppl 4, pp. D49–53, Sept. 2008.

[7] M. Krystal, R. M. Elliott, E. W. Benz, J. F. Young, and P. Palese, "Evolution of influenza A and B viruses: conservation of structural features in the hemagglutinin genes," *Proceedings of the National Academy of Sciences*, vol. 79, pp. 4800–4804, Aug. 1982.

[8] Centre for Disease Control and Prevention, "CDC - Influenza," 2014.

[9] A. revision of the system of nomenclature for influenza viruses, "A WHO memorandum," *Bulletin of the World Health Organization*, vol. 58, pp. 585–91, Jan. 1980.

[10] R. Lamb and R. Kurg, "Orthomyxoviridae: the viruses and their replication," in *Fields Virology* (D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Strauss, eds.), pp. 1487–1531, Lippincott Williams & Wilkins, 4 ed., 2001.

[11] W. Chen, P. A. Calvo, D. Malide, J. Gibbs, U. Schubert, I. Bacik, S. Basta, R. O'Neill, J. Schickli, P. Palese, P. Henklein, J. R. Bennink, and J. W. Yewdell, "A novel influenza A virus mitochondrial protein that induces cell death," *Nature medicine*, vol. 7, pp. 1306–12, Dec. 2001.

[12] H. M. Wise, A. Foeglein, J. Sun, R. M. Dalton, S. Patel, W. Howard, E. C. Anderson, W. S. Barclay, and P. Digard, "A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA," *Journal of virology*, vol. 83, pp. 8021–31, Aug. 2009.

[13] S. Mubareka, A. C. Lowen, J. Steel, A. L. Coates, A. Garcia Sastre, and P. Palese, "Transmission of Influenza Virus via Aerosols and Fomites in the Guinea Pig Model," *The Journal of Infectious Diseases*, vol. 199, pp. 858–865, Mar. 2009.

[14] P. Fabian, J. J. McDevitt, W. H. DeHaan, R. O. P. Fung, B. J. Cowling, K. H. Chan, G. M. Leung, and D. K. Milton, "Influenza virus in human exhaled breath: an observational study," *PloS one*, vol. 3, p. e2691, Jan. 2008.

[15] K. L. Hartshorn, "New look at an old problem: bacterial superinfection after influenza," *The American journal of pathology*, vol. 176, pp. 536–9, Feb. 2010.

[16] M. Tashiro, P. Ciborowski, H. D. Klenk, G. Pulverer, and R. Rott, "Role of Staphylococcus protease in the development of influenza pneumonia," *Nature*, vol. 325, no. 6104, pp. 536–7.

[17] K. D. Patterson and G. F. Pyle, "The geography and mortality of the 1918 influenza pandemic," *Bulletin of the history of medicine*, vol. 65, pp. 4–21, Jan. 1991.

[18] J. Greene and K. Moline, *The Bird Flu Pandemic*. St. Martin's Press, 2006.

[19] E. D. Kilbourne, "Influenza pandemics of the 20th century," *Emerging infectious diseases*, vol. 12, pp. 9–14, Jan. 2006.

[20] C. E. Mills, J. M. Robins, and M. Lipsitch, "Transmissibility of 1918 pandemic influenza," *Nature*, vol. 432, pp. 904–6, Dec. 2004.

[21] K. B. Walsh, J. R. Teijaro, P. R. Wilker, A. Jatzek, D. M. Fremgen, S. C. Das, T. Watanabe, M. Hatta, K. Shinya, M. Suresh, Y. Kawaoka, H. Rosen, and M. B. A. Oldstone, "Suppression of cytokine storm with a sphingosine analog provides protection against pathogenic influenza virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 12018–23, July 2011.

[22] H. Manchanda, N. Seidel, A. Krumbholz, A. Sauerbrei, M. Schmidtke, and R. Guthke, "Within-host influenza dynamics: a small-scale mathematical modeling approach," *Bio Systems*, vol. 118, pp. 51–9, Apr. 2014.

[23] S. Jin, Y. Li, R. Pan, and X. Zou, "Characterizing and controlling the inflammatory network during influenza A virus infection," *Scientific reports*, vol. 4, p. 3799, Jan. 2014.

[24] E. W. Larson, J. W. Dominik, A. H. Rowberg, and G. A. Higbee, "Influenza virus population dynamics in the respiratory tract of experimentally infected mice," *Infection and immunity*, vol. 13, pp. 438–47, Feb. 1976.

[25] A. S. Perelson, D. E. Kirschner, and R. De Boer, "Dynamics of HIV infection of CD4+ T cells," *Mathematical biosciences*, vol. 114, pp. 81–125, Mar. 1993.

[26] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, "HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time," *Science (New York, N.Y.)*, vol. 271, pp. 1582–6, Mar. 1996.

[27] A. S. Perelson, "Modelling viral and immune system dynamics," *Nature reviews. Immunology*, vol. 2, pp. 28–36, Jan. 2002.

[28] P. Baccam, C. Beauchemin, C. A. Macken, F. G. Hayden, and A. S. Perelson, "Kinetics of influenza A virus infection in humans," *Journal of virology*, vol. 80, pp. 7590–9, Aug. 2006.

[29] C. A. A. Beauchemin, J. J. McSharry, G. L. Drusano, J. T. Nguyen, G. T. Went, R. M. Ribeiro, and A. S. Perelson, "Modeling amantadine treatment of influenza A virus in vitro," *Journal of theoretical biology*, vol. 254, pp. 439–51, Sept. 2008.

[30] A. M. Smith, F. R. Adler, and A. S. Perelson, "An accurate two-phase approximate solution to an acute viral infection model," *Journal of mathematical biology*, vol. 60, pp. 711–26, May 2010.

[31] G. A. Bocharov and A. A. Romanyukha, "Mathematical model of antiviral immune response. III. Influenza A virus infection.," *Journal of theoretical biology*, vol. 167, pp. 323–60, Apr. 1994.

[32] R. Brun, P. Reichert, and H. R. Künsch, "Practical identifiability analysis of large environmental simulation models," *Water Resources Research*, vol. 37, pp. 1015–1030, Apr. 2001.

[33] H. Trottier and P. Philippe, "Deterministic Modeling Of Infectious Diseases: Theory And Methods," *The Internet Journal of Infectious Diseases.*, vol. 1, no. 2, 2000.

[34] C. I. Siettos and L. Russo, "Mathematical modeling of infectious disease dynamics," *Virulence*, vol. 4, pp. 295–306, May 2013.

[35] B. P. Holder and C. A. A. Beauchemin, "Exploring the effect of biological delays in kinetic models of influenza within a host or cell culture," *BMC public health*, vol. 11 Suppl 1, p. S10, Jan. 2011.

[36] A. M. Smith and A. S. Perelson, "Influenza A virus infection kinetics: quantitative data and models," *Wiley interdisciplinary reviews. Systems biology and medicine*, vol. 3, no. 4, pp. 429–45.

[37] H. Wu, P. Ruan, A. A. Ding, J. L. Sullivan, and K. Luzuriaga, "Inappropriate model-fitting methods may lead to significant underestimates of viral decay rates in HIV dynamic studies," *Journal of acquired immune deficiency syndromes (1999)*, vol. 21, pp. 426–8, Aug. 1999.

[38] M. Sanayei, B. Arya, E. M. Santini, and S. Wadia-Fascetti, "Significance of Modeling Error in Structural Parameter Estimation," *Computer-Aided Civil and Infrastructure Engineering*, vol. 16, pp. 12–27, Jan. 2001.

[39] B. Hancioglu, D. Swigon, and G. Clermont, "A dynamical model of human immune response to influenza A virus infection," *Journal of theoretical biology*, vol. 246, pp. 70–86, May 2007.

[40] A. Handel, I. M. Longini, and R. Antia, "Towards a quantitative understanding of the within-host dynamics of influenza A infections," *Journal of the Royal Society, Interface / the Royal Society*, vol. 7, pp. 35–47, Jan. 2010.

[41] H. Miao, J. A. Hollenbaugh, M. S. Zand, J. Holden-Wiltse, T. R. Mosmann, A. S. Perelson, H. Wu, and D. J. Topham, "Quantifying the early immune response and adaptive immune response kinetics in mice infected with influenza A virus," *Journal of virology*, vol. 84, pp. 6687–98, July 2010.

[42] K. A. Pawelek, G. T. Huynh, M. Quinlivan, A. Cullinane, L. Rong, and A. S. Perelson, "Modeling within-host dynamics of influenza virus infection including immune responses," *PLoS computational biology*, vol. 8, p. e1002588, Jan. 2012.

[43] M. Castro-Melchor, S. Charaniya, G. Karypis, E. Takano, and W.-S. Hu, "Genome-wide inference of regulatory networks in Streptomyces coelicolor," *BMC genomics*, vol. 11, p. 578, Jan. 2010.

[44] M. Morita, K. Kuba, A. Ichikawa, M. Nakayama, J. Katahira, R. Iwamoto, T. Watanebe, S. Sakabe, T. Daidoji, S. Nakamura, A. Kadowaki, T. Ohto, H. Nakanishi, R. Taguchi, T. Nakaya, M. Murakami, Y. Yoneda, H. Arai,

Y. Kawaoka, J. M. Penninger, M. Arita, and Y. Imai, "The lipid mediator protectin D1 inhibits influenza virus replication and improves severe influenza," *Cell*, vol. 153, pp. 112–25, Mar. 2013.

[45] C. Cillóniz, K. Shinya, X. Peng, M. J. Korth, S. C. Proll, L. D. Aicher, V. S. Carter, J. H. Chang, D. Kobasa, F. Feldmann, J. E. Strong, H. Feldmann, Y. Kawaoka, and M. G. Katze, "Lethal influenza virus infection in macaques is associated with early dysregulation of inflammatory related genes," *PLoS pathogens*, vol. 5, p. e1000604, Oct. 2009.

[46] R. Kumar, G. Clermont, Y. Vodovotz, and C. C. Chow, "The dynamics of acute inflammation.," *Journal of theoretical biology*, vol. 230, pp. 145–55, Sept. 2004.

[47] A. M. Smith, F. R. Adler, J. L. McAuley, R. N. Gutenkunst, R. M. Ribeiro, J. A. McCullers, and A. S. Perelson, "Effect of 1918 PB1-F2 expression on influenza A virus infection kinetics," *PLoS computational biology*, vol. 7, p. e1001081, Feb. 2011.

[48] L. Canini and F. Carrat, "Population modeling of influenza A/H1N1 virus kinetics and symptom dynamics," *Journal of virology*, vol. 85, pp. 2764–70, Mar. 2011.

[49] R. A. Saenz, M. Quinlivan, D. Elton, S. Macrae, A. S. Blunden, J. A. Mumford, J. M. Daly, P. Digard, A. Cullinane, B. T. Grenfell, J. W. McCauley, J. L. N. Wood, and J. R. Gog, "Dynamics of influenza virus infection and pathology," *Journal of virology*, vol. 84, pp. 3974–83, Apr. 2010.

[50] H. Y. Lee, D. J. Topham, S. Y. Park, J. Hollenbaugh, J. Treanor, T. R. Mosmann, X. Jin, B. M. Ward, H. Miao, J. Holden-Wiltse, A. S. Perelson, M. Zand, and H. Wu, "Simulation and prediction of the adaptive immune response to influenza A virus infection," *Journal of virology*, vol. 83, pp. 7151–65, July 2009.

[51] K. A. Frazer, "Decoding the human genome," *Genome research*, vol. 22, pp. 1599–601, Sept. 2012.

[52] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science (New York, N.Y.)*, vol. 270, pp. 467–70, Oct. 1995.

[53] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature biotechnology*, vol. 14, pp. 1675–80, Dec. 1996.

[54] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature reviews. Genetics*, vol. 7, pp. 55–65, Jan. 2006.

[55] J.-B. Fan, K. L. Gunderson, M. Bibikova, J. M. Yeakley, J. Chen, E. Wickham Garcia, L. L. Lebruska, M. Laurent, R. Shen, and D. Barker, "Illumina universal bead arrays," *Methods in enzymology*, vol. 410, pp. 57–73, Jan. 2006.

[56] K. Kuhn, S. C. Baker, E. Chudin, M.-H. Lieu, S. Oeser, H. Bennett, P. Rigault, D. Barker, T. K. McDaniel, and M. S. Chee, "A novel, high-performance random array platform for quantitative gene expression profiling ,"

[57] K. L. Gunderson, S. Kruglyak, M. S. Graige, F. Garcia, B. G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, D. Doucet, M. Milewski, R. Yang, C. Siegmund, J. Haas, L. Zhou, A. Oliphant, J.-B. Fan, S. Barnard, and M. S. Chee, "Decoding randomly ordered DNA arrays," *Genome research*, vol. 14, pp. 870–7, May 2004.

[58] K. Ljungberg, A. McBrayer, J. V. Camp, Y.-K. Chu, R. Tapp, D. L. Noah, S. Grimes, M. L. Proctor, P. Liljeström, C. B. Jonsson, and C. E. Bruder, "Host gene expression signatures discriminate between ferrets infected with genetically similar H1N1 strains," *PloS one*, vol. 7, p. e40743, Jan. 2012.

[59] C. Pommerenke, E. Wilk, B. Srivastava, A. Schulze, N. Novoselova, R. Geffers, and K. Schughart, "Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses," *PloS one*, vol. 7, p. e41169, Jan. 2012.

[60] P. S. Askovich, C. J. Sanders, C. M. Rosenberger, A. H. Diercks, P. Dash, G. Navarro, P. Vogel, P. C. Doherty, P. G. Thomas, and A. Aderem, "Differential host response, rather than early viral replication efficiency, correlates with pathogenicity caused by influenza viruses," *PloS one*, vol. 8, p. e74863, Jan. 2013.

[61] M. Ding, L. Lu, and L. A. Toth, "Gene expression in lung and basal forebrain during influenza infection in mice," *Genes, brain, and behavior*, vol. 7, pp. 173–83, Mar. 2008.

[62] H. Kollmus, E. Wilk, and K. Schughart, "Systems biology and systems genetics - novel innovative approaches to study host-pathogen interactions during influenza infection," *Current opinion in virology*, vol. 6, pp. 47–54, June 2014.

[63] C. Schrader and J. Süss, "Genetic characterization of a porcine H1N2 influenza virus strain isolated in Germany," *Intervirology*, vol. 46, pp. 66–70, Jan. 2003.

[64] K. Bauer, R. Dürrwald, M. Schlegel, K. Pfarr, D. Topf, N. Wiesener, H.-M. Dahse, P. Wutzler, and M. Schmidtke, "Neuraminidase inhibitor susceptibility of swine influenza A viruses isolated in Germany between 1981 and 2008," *Medical microbiology and immunology*, vol. 201, pp. 61–72, Feb. 2012.

[65] E. Wattrang, D. M. Jessett, P. Yates, L. Fuxler, and D. Hannant, "Experimental infection of ponies with equine influenza A2 (H3N8) virus strains of different pathogenicity elicits varying interferon and interleukin-6 responses," *Viral immunology*, vol. 16, pp. 57–67, Jan. 2003.

[66] R. Goris and A. Jan, "Multiple-Organ Failure," *Archives of Surgery*, vol. 120, p. 1109, Oct. 1985.

[67] A. TAKALA, I. JOUSELA, S. JANSSON, K. OLKKOLA, O. TAKKUNEN, A. ORPANA, S. KARONEN, and H. REPO, "Markers of systemic inflammation predicting organ failure in community-acquired septic shock," Nov. 1999.

[68] K. Soetaert, T. Petzoldt, and R. Setzer, "Solving Differential Equations in R: Package deSolve," *Journal of Statistical Software*, vol. 33, pp. 1–25, 2010.

[69] K. Soetaert and T. Petzoldt, "Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME," July 2010.

[70] J. Moré Jorge, "The Levenberg-Marquardt algorithm: Implementation and theory," *Lecture Notes in Mathematics*, vol. 630, pp. 105–116, 1978.

[71] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1923–9, Aug. 2009.

[72] H. Haario, M. Laine, A. Mira, and E. Saksman, "DRAM: Efficient adaptive MCMC," *Statistics and Computing*, vol. 16, pp. 339–354, Dec. 2006.

[73] L. Tierney, H. Doss, J. Besag, C. P. Robert, K. S. Chan, and C. J. Geyer, "Markov chains for exploring posterior distributions," *Annals of statistics*, vol. 22, no. 4, pp. 1701–1762.

[74] G. Kuczera and E. Parent, "Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm," *Journal of Hydrology*, vol. 211, pp. 69–85, Nov. 1998.

[75] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*. MIT Press, 2001.

[76] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, vol. 21, p. 1087, Dec. 1953.

[77] P. Du, W. A. Kibbe, and S. M. Lin, "lumi: a pipeline for processing Illumina microarray," *Bioinformatics (Oxford, England)*, vol. 24, pp. 1547–8, July 2008.

[78] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe, "Model-based variance-stabilizing transformation for Illumina microarray data," *Nucleic acids research*, vol. 36, p. e11, Feb. 2008.

[79] T. J. Hardcastle and K. A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC bioinformatics*, vol. 11, p. 422, Jan. 2010.

[80] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. 57, no. 1, pp. 289–300, 1995.

[81] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, The Regents of the University of California, 1967.

[82] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, pp. 32–57, Jan. 1973.

[83] J. C. Bezdek, Robert Ehrlich, and William Full, "FCM - The Fuzzy c-Means Clustering Algorithm," *Computers and Geosciences*, vol. 10, pp. 191–203, 1984.

[84] R. Guthke, U. Möller, M. Hoffmann, F. Thies, and S. Töpfer, "Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection," *Bioinformatics (Oxford, England)*, vol. 21, pp. 1626–34, Apr. 2005.

[85] S. Toepfer, R. Guthke, D. Driesch, D. Woetzel, and M. Pfaff, "The netgenerator algorithm: reconstruction of gene regulatory networks," pp. 119–130, May 2006.

[86] J. Linde, D. Wilson, B. Hube, and R. Guthke, "Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells," *BMC systems biology*, vol. 4, p. 148, Jan. 2010.

[87] M. Weber, S. G. Henkel, S. Vlaic, R. Guthke, E. J. van Zoelen, and D. Driesch, "Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0," *BMC systems biology*, vol. 7, p. 1, Jan. 2013.

[88] Kilander A, Rykkvin R, Dudman SG, and Hungnes O, "Observed association between the HA1 mutation D222G in the 2009 pandemic influenza A(H1N1) virus and severe clinical outcome, Norway 2009-2010," *Eurosurveillance*, vol. 15, Apr. 2010.

[89] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, p. Article3, Jan. 2004.

[90] R Development Core Team, "R: a language and environment for statistical computing," Vienna, Austria: R Foundation for Statistical Computing, 2011.

[91] B. G. Hale, R. A. Albrecht, and A. García-Sastre, "Innate immune evasion strategies of influenza viruses," *Future microbiology*, vol. 5, pp. 23–41, Jan. 2010.

[92] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature protocols*, vol. 4, pp. 44–57, Jan. 2009.

[93] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, pp. 1–13, Jan. 2009.

[94] X. Hu, S. D. Chakravarty, and L. B. Ivashkiv, "Regulation of interferon and Toll-like receptor signaling during macrophage activation by opposing feedforward and feedback inhibition mechanisms," *Immunological reviews*, vol. 226, pp. 41–56, Dec. 2008.

[95] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio–the analysis and navigation of molecular networks," *Bioinformatics (Oxford, England)*, vol. 19, pp. 2155–7, Nov. 2003.

[96] J. Linde, P. Hortschansky, E. Fazius, A. A. Brakhage, R. Guthke, and H. Haas, "Regulatory interactions for iron homeostasis in Aspergillus fumigatus inferred by a Systems Biology approach," *BMC systems biology*, vol. 6, p. 6, Jan. 2012.

[97] M. Weber, A. M. Sotoca, P. Kupfer, R. Guthke, and E. J. van Zoelen, "Dynamic modelling of microRNA regulation during mesenchymal stem cell differentiation," *BMC systems biology*, vol. 7, p. 124, Jan. 2013.

[98] S. Ramachandra, J. Linde, M. Brock, R. Guthke, B. Hube, and S. Brunke, "Regulatory networks controlling nitrogen sensing and uptake in Candida albicans," *PloS one*, vol. 9, p. e92734, Jan. 2014.

[99] L. Tierney, J. Linde, S. Müller, S. Brunke, J. C. Molina, B. Hube, U. Schöck, R. Guthke, and K. Kuchler, "An Interspecies Regulatory Network Inferred from Simultaneous RNA-seq of Candida albicans Invading Innate Immune Cells," *Frontiers in microbiology*, vol. 3, p. 85, Jan. 2012.

[100] K. Breuer, A. K. Foroushani, M. R. Laird, C. Chen, A. Sribnaia, R. Lo, G. L. Winsor, R. E. W. Hancock, F. S. L. Brinkman, and D. J. Lynn, "InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation," *Nucleic acids research*, vol. 41, pp. D1228–33, Jan. 2013.

[101] H. X. Zhang, Z. X. Liu, Y. P. Sun, J. Zhu, S. Y. Lu, X. S. Liu, Q. H. Huang, Y. Y. Xie, H. B. Zhu, S. Y. Dang, H. F. Chen, G. Y. Zheng, Y. X. Li, Y. Kuang, J. Fei, S. J. Chen, Z. Chen, and Z. G. Wang, "Rig-I regulates NF-κB activity through binding to Nf-κb1 3'-UTR mRNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 6459–64, Apr. 2013.

[102] H. Zeng, C. Goldsmith, P. Thawatsupha, M. Chittaganpitch, S. Waicharoen, S. Zaki, T. M. Tumpey, and J. M. Katz, "Highly pathogenic avian influenza H5N1 viruses elicit an attenuated type i interferon response in polarized human bronchial epithelial cells," *Journal of virology*, vol. 81, pp. 12439–49, Nov. 2007.

[103] Z. Guo, L.-m. Chen, H. Zeng, J. A. Gomez, J. Plowden, T. Fujita, J. M. Katz, R. O. Donis, and S. Sambhara, "NS1 protein of influenza A virus inhibits the function of intracytoplasmic pathogen sensor, RIG-I," *American journal of respiratory cell and molecular biology*, vol. 36, pp. 263–9, Mar. 2007.

[104] K. V. Chakravarthy, A. C. Bonoiu, W. G. Davis, P. Ranjan, H. Ding, R. Hu, J. B. Bowzard, E. J. Bergey, J. M. Katz, P. R. Knight, S. Sambhara, and P. N. Prasad, "Gold nanorod delivery of an ssRNA immune activator inhibits pandemic H1N1 influenza viral replication," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 10172–7, June 2010.

[105] M. Pitchaimani, C. Monica, and M. Divya, "Stability analysis for HIV infection delay model with protease inhibitor," *Bio Systems*, vol. 114, pp. 118–24, Nov. 2013.

[106] J. A. Wiley, A. Cerwenka, J. R. Harkema, R. W. Dutton, and A. G. Harmsen, "Production of interferon-gamma by influenza hemagglutinin-specific CD8 effector

T cells influences the development of pulmonary immunopathology," *The American journal of pathology*, vol. 158, pp. 119–30, Jan. 2001.

[107] T. Ostler, W. Davidson, and S. Ehl, "Virus clearance and immunopathology by CD8(+) T cells during infection with respiratory syncytial virus are mediated by IFN-gamma," *European journal of immunology*, vol. 32, pp. 2117–23, Aug. 2002.

[108] M. Brandes, F. Klauschen, S. Kuchen, and R. N. Germain, "A systems analysis identifies a feedforward inflammatory circuit leading to lethal influenza infection," *Cell*, vol. 154, pp. 197–212, July 2013.

[109] I. Rahman, "Oxidative stress, chromatin remodeling and gene transcription in inflammation and chronic lung diseases," *Journal of biochemistry and molecular biology*, vol. 36, pp. 95–109, Jan. 2003.

[110] D. Kobasa, A. Takada, K. Shinya, M. Hatta, P. Halfmann, S. Theriault, H. Suzuki, H. Nishimura, K. Mitamura, N. Sugaya, T. Usui, T. Murata, Y. Maeda, S. Watanabe, M. Suresh, T. Suzuki, Y. Suzuki, H. Feldmann, and Y. Kawaoka, "Enhanced virulence of influenza A viruses with the haemagglutinin of the 1918 pandemic virus," *Nature*, vol. 431, pp. 703–7, Oct. 2004.

[111] T. M. Tumpey, A. García-Sastre, J. K. Taubenberger, P. Palese, D. E. Swayne, M. J. Pantin-Jackwood, S. Schultz-Cherry, A. Solórzano, N. Van Rooijen, J. M. Katz, and C. F. Basler, "Pathogenicity of influenza viruses with genes from the 1918 pandemic virus: functional roles of alveolar macrophages and neutrophils in limiting virus replication and mortality in mice," *Journal of virology*, vol. 79, pp. 14933–44, Dec. 2005.

Declaration

I confirm that I'm familiar with the relevant course of examination for doctoral candidates (Promotionsordnung).

I have composed and written the dissertation by myself and I have acknowledged all additional assistance, personal communication, and sources within the work.

I have not enlisted the assistance of a doctoral consultant (Promotionsberater) and no third parties have received either direct or indirect monetary benefits from me for work connected to the submitted dissertation.

I have not submitted the dissertation as an examination paper for a state or other scientific examination.

Signed: _____

Date: _____

Himanshu Manchanda

Curriculum Vitae

Institute for Bioinformatics,
Uni. Medicine, Greifswald, Germany
(+49) 176 5692 6596
bioinfo.himanshu@gmail.com

Research Interest

My research focuses lie in the systems biology of an infection process. Particularly, I am interested in high throughput data analysis, in particular reconstruction of the gene regulatory network. I am interested in analysing how different influenza strain-based on their pathogenicity, interact with host and regulate the host immune system especially the inflammatory response. During my PhD, I have applied and studied several Ordinary Differential Equation (ODE) based mathematical models of influenza kinetics which helps in quantifying different viral strains based on their pathogenicity. Apart from that, I have further developed expertise in high throughput data analysis, especially microarray data and construction of gene regulatory networks.

Research Experience

Oct, 2011 - Dec, 2011 **Project Assistant**, National Brain Research Centre (NBRC), India.
Under Dr Ranjit Giri; I was involved in proteomics data analysis but unfortunately could not work for longer as I got the PhD offer from HKI, Jena.

May, 2011 - Aug, 2011 **Project Assistant (Part-time)**, School of Life Sciences, Jawaharlal Nehru University (JNU), New Delhi, India.
Under Prof. Ashwani Pareek; The work involves the micro array data analysis of plant's response to various abiotic stresses, so as to exploit this knowledge to improve crop yield, quality and quantity under adverse enviromental conditions.

Sep, 2009 - Mar, 2011 **Junior Research Fellow**, International Centre for Genetic Engineering and Biotechnology (ICGEB), New Delhi, India.
Project funded by PepsiCo under the supervision of Dr. Kanury Venkata Subba Rao. I worked as a data analyst. My work here involved the study and analysis of the expression level of mRNAs of the whole Genome of normal and Diabetic mice, using high throughput techniques based on DNA microarray and Proteome (MALDI-TOF) technology. The aim is to develop Genetic Regulatory Networks and determine the fingerprints or Genetic signatures through Machine learning (PLSR) and Protein-Protein interactions for the early detection of Diabetes.

Jan, 2009 - July, 2009 **Research Intern**, Institute of Bioinformatics and Applied Biotechnology (IBAB), Bangalore, India.
Under Dr. Kshitish Acharya; My work here involved the data curation and data analysis related to tissue and disease specific database. The objective of the work was to arrive at better set of gene signatures for the disease endometriosis. Approach involved collection of all gene expression datasets, across multiple studies and methods; compare the data; derive a reliability score; analyze and identifying novel gene signature lists.

May, 2009 - June, 2009 **Lectures Taken**, One month training on User End Bioinformatics held at IBAB, Bangalore, India.
Actively contributed to two important training sessions during the one-month workshop, with extensive hands on training:
1) Basics of algorithm involved in sequence analysis.
2) Introduction to drug designing and computational modelling of molecules.

Education

2015- **Post Doctoral Scientist, Bioinformatics**, Institute for Bioinformatics, Greifswald, Germany.
Currently working with Prof. Lars Kaderali and involved in mutliple projects dealing with high throughput data analysis to bayesian modeling of epidemiological studies involving the Study of Health in Pomerania (SHIP).

2012- 2015 **Ph.D. Candidate, Bioinformatics**, HKI, Jena, Germany.
Supervisor: Prof. Reinhard Guthke, Thesis title: *"Understanding the pathogenesis of severe influenza A infection by using Mathematical modelling, and High throughput data analysis approaches"*.

2007-2009 **M.Sc.Bioinformatics**, Guru Nanak Dev University (GNDU), India.
Supervisor: Prof. Kshitish Acharya, Thesis title: *"Deriving signature set of genes for endometriosis, using a newly created database of genome-wide expression profiling data"*.
University Rank 2

2004-2007 **B.Sc.Biotechnology**, DAV College, GNDU, Amritsar, India.
College Rank 5.

Publications

Manchanda H, Seidel N, Krumbholz A, Sauerbrei A, Schmidtke M, Guthke R. Within-host influenza dynamics: A small-scale mathematical modelling approach. *Biosystems* 2014, 118:51-59.

Manchanda H, Seidel N, Markus F. Blaess, Ralf A. Cluas, Jörg Linde, Krumbholz A, Sauerbrei A, Schmidtke M, Guthke R. A positive feedback-loop regulates the transcriptional response to murine pandemic influenza infection. ***under submission.***

Felix Bormann1, Manuel Rodríguez-Paredes,Sabine Hagemann, **Himanshu Manchanda**, Boris Kristof, Julian Gutekunst, Rainer Haas, Lara Terstegen, Horst Wenck, Lars Kaderali, Marc Winnefeld, and Frank Lyko. Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. ***under submission.***

Conference Participation / Seminar

Manchanda H, "Mathematical modelling of pandemic influenza A virus-induced disease in mice at Influenza 2013: One Influenza, One World, One Health Conference, Oxford, UK, on 16-19, September 2013. **Oral presentation**

Manchanda H, "Mathematical modelling of pandemic influenza A virus-induced disease in mice at Influenza 2013: One Influenza, One World, One Health Conference, Oxford, UK, on 16-19, September 2013. **Poster presentation**

Manchanda H, "Modelling of influenza virus infection in mice at German Conference on Bioinformatics (GCB), Jena, Germany, on 19-22, September 2012. **Poster presentation**