IDENTIFYING STRUCTURAL CHANGES IN EVOLVING VIRAL PROTEINS

A Dissertation in
Integrative Biosciences

by
Sunando Roy

© 2011 Sunando Roy

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2011
The dissertation of Sunando Roy was reviewed and approved* by the following:

Mary Poss
Professor of Biology and in Veterinary and Biomedical Sciences
Dissertation Advisor
Chair of Committee

Francesca Chiaromonte
Professor of Statistics and Public Health Sciences

Ross C. Hardison
T. Ming Chu Professor of Biochemistry and Molecular Biology

Philip C. Bevilacqua
Professor of Chemistry

Peter Hudson
Willaman Professor of Biology
Director, Huck Institutes of the Life Sciences

*Signatures are on file in the Graduate School
Proteins are macromolecules that facilitate the basic biochemical functions of a cell. Proteins help organisms adapt to their new environment by changing their function, which in turn is often accompanied by changes in structure. When a virus moves from one host to another, the move is accompanied by changes in the sequence of the virus. These changes are studied using phylogenetic tools that identify sites in the genome, which are important for adaptation in the new host (positively selected sites). Studying how these changes in the genome sequence affects change in the structure of viral proteins is still very challenging, as a one-to-one relationship does not exist between changes in sequence and changes in structure. The primary structure of the protein is its amino acid sequence that determines the final tertiary structure of the protein. The rules that govern the folding of proteins starting from the linear amino acid sequence are still poorly understood. Over the last five decades progress has been made in predicting the secondary structure of proteins using robust computational tools. The secondary structure, which is an intermediate step (alpha helix, beta sheet, and coils) between primary and tertiary structure, is a property determined by a combination of amino acids. We have used six predicted secondary structural properties to construct a multivariate statistical pipeline that identifies differences in structure at amino acid positions between groups of related viral proteins (Human and Avian Metapneumovirus, HIV and SIV). This pipeline when used with phylogenetic analysis of linear sequence data adds value by predicting structural changes that arise in the course of viral adaptation to new host.
# TABLE OF CONTENTS

List of Figures .................................................................................................................. vii

List of Tables ...................................................................................................................... viii

Acknowledgements .......................................................................................................... ix

Chapter 1: Introduction .................................................................................................... 1
  1.1 Background ............................................................................................................... 1
  1.2 Objectives and Outline ............................................................................................ 7

Chapter 2: Multivariate Statistical Analyses Demonstrate Unique Host Immune Responses to Single and Dual Lentiviral Infection ........................................................................ 11
  2.1 Introduction ............................................................................................................. 11
  2.2 Materials and Methods ........................................................................................... 14
    2.2.1 Study Design and Pre-processing .................................................................... 14
    2.2.2 Statistical Analyses ......................................................................................... 16
    2.2.3 Principal Component Analysis ...................................................................... 17
    2.2.4 Multivariate Analysis of Variance .................................................................. 18
    2.2.5 Linear Discriminant Analysis ....................................................................... 19
  2.3 Results ....................................................................................................................... 21
    2.3.1 Immune response parameters co-vary ........................................................... 21
    2.3.2 Mean immune responses differ significantly across infection status ............ 25
    2.3.3 Specific immune profiles, which evolve over time, discriminate among infection status ........................................................................................................... 28
  2.4 Discussion ................................................................................................................ 32

Chapter 3: Developing a statistical pipeline to identify secondary structure changes between related groups of proteins ......................................................................................... 38
  3.1 Introduction ............................................................................................................. 38
  3.2 Material and Methods ............................................................................................ 42
    3.2.1 Sequence Data ................................................................................................. 42
    3.2.2 Prediction of secondary structure .................................................................... 43
    3.2.3 Within group and between group sequence variability ................................... 44
    3.2.4 Identification of structurally informative sites (SInS) ...................................... 46
    3.2.5 Coefficient of variation (CV) ........................................................................ 48
  3.3 Results ....................................................................................................................... 49
    3.3.1 Prediction servers and SInS .......................................................................... 49
LIST OF FIGURES

Figure 2-1: (A) Box-plots of (absolute) pair-wise correlations and (B) Scree plots from PCA..........................23

Figure 2-2: Bi-plots from LDA, for days 31, 37, 52, and 59.................................................................30

Figure 3-1: Structures of the Ga and the Gb domain of the G protein of Streptococcus with difference percentages of sequence identity.................................................................43

Figure 3-2: The distances between Ga and Gb for the five structure prediction servers...........50

Figure 3-3: The distances between Ga and Gb for values of between group variability...........52

Figure 3-4: The CV (Coefficient of variation) plotted against distance between Ga and Gb (A) Distance from JPred, (B) Distance from A1B1.................................................................55

Figure 4-1: Structurally informative sites (SInS) across the length of the membrane associated glycoproteins F (A) and G (B)......................................................................................68

Figure 4-2: Relationship between percentages of polymorphic sites and (A) phylogenetically informative sites (PInS), and (B) between percentages of structurally informative sites (SInS)..............................................................................................................70

Figure 4-3: Neighbor joining trees showing the phylogenetic relationships among different subtypes of metapneumovirus isolated from human and avian hosts. (A) based on the concatenated protein sequences, (B) on the concatenated SInS for the whole genome.......71

Figure 5-1: Structurally informative sites (SInS) across the length of the Reverse Transcriptase protein. (A) Between HIV1/SIVcpz and (B) Between HIV1/HIV2.................................................84

Figure 5-2: SInS highlighted on the crystal structure of HIV1 RT (1RTJ) (A) Front (B) Back. The p66 subunit is colored blue whereas the p51 is shown in red..................................................88

Figure 5-3: Bi-plots from LDA, for the four primate lentiviral groups (HIV1, SIVcpz, HIV2 and SIVsm)..........................................................................................................................90
LIST OF TABLES

Table 2-1: List of immunological response parameters used in the analysis……………………..15

Table 2-2: Pearson's coefficients of the ten strongest (absolute) pair-wise correlations between immunological responses for each day of analysis.................................................................24

Table 2-3: Spearman's coefficients of the ten strongest (absolute) pair-wise correlations between immunological responses for each day of analysis.................................................................25

Table 2-4: Estimated effects (E) and Pillai's trace p-values (p) for a two-way MANOVA scheme comprising simple and interaction effects of PLV and FIVC, for each day of analysis………26

Table 2-5: Estimated effects (E) and Bonferroni corrected F-based p-values (p) for two-way ANOVA schemes applied to each immunological parameter separately, for each day of analysis…………………………………………………………………………………………27

Table 2-6: Total variances within infection status groups on different days.........................32

Table 3-1: Between group and within group sequence variability of 4 groups of Ga (A1 – A4) and 3 groups of Gb (B1 – B3) sequences……………………………………………………………………………46

Table 3-2: Sensitivity (no. of identified SInS over the length of the protein) and Specificity (no. of correctly identified SInS over no. of true SInS) values for each prediction server………50

Table 3-3: Sensitivity (no. of identified SInS over the length of the protein) and Specificity (no. of correctly identified SInS over no. of true SInS) values for different between group variability……………………………………………………………………………………53

Table 4-1: Evolutionary genetics and structurally informative sites of the proteins in metapneumovirus..............................................................................................................................66

Table 5-1: Number of SInS between primate lentivirus groups........................................83

Table 5-2: Percentage of SInS distributed in different domains of RT................................85

Table 5-3: Percentage of secondary structural properties that differ at SInS.......................86
ACKNOWLEDGEMENTS

Five and a half years later without any warning it’s over and in a moment I would have achieved what I set out for in Aug 2006, a PhD. It has been a long winding road on a journey that at many moments felt without a definite destination. Well it is finally over and it would not have been possible without my support system in this place I called home over the last few years. I would like to thank my advisor Dr. Mary Poss for her support over the years along with a wonderful committee comprising of Dr. Francesca Chiaromonte, Dr. Ross Hardison, and Dr. Philip Bevilacqua. I would also thank my labmates past and present especially Yeeling, Abinash, Daniel, and Oekyung for their wonderful support and interesting discussion about the ways of the PhD. I would like to thank some wonderful people I met whom I am proud to call my friends. Rohan, Kokonad, Prabhasa, Gaurav, Amit, and Swayamjit thank you for sharing this fun journey with me, without your individual contributions (too long to list) the journey would just be confined to my research, imagine that horror©. I would also like to thank my parents for supporting my career choices and knowing very well that I will be borrowing money from them continuously over the next 100 years. Last but not the least I would like to thank Atisha for being a close part of my life for the last three years and I will be looking forward to our time together for the years to come.

Now that it is over there is one thing I have realized, this journey was not about the destination it was about survival and I SURVIVED!!!!!!!!!!
Chapter 1

Introduction

1.1 Background

Proteins are essential macromolecules that facilitate most functions of an organism. The function of any protein is strongly determined by its three dimensional structure, which in turn is determined by its linear amino acid sequence [1,2,3]. It is a challenge still to identify completely all the rules that govern folding of the linear sequence to its three dimensional structures. Proteins evolve by substitutions in their sequences which in turn affect their structure and function, although most substitutions that arise are neutral relative to changes in structure [4]. Some proteins can accommodate amino acid substitutions more than others without changing their overall structure and function [4]. Thus, no one-to-one relationship exists between changes in sequence and structure and it is still an open question in biology to understand completely how changes in sequence affect the structure and function of proteins [5,6,7].

There are currently multiple ways to address the question of how change in sequence affects change in structure. The first way is to compare the solved protein structures (X-ray, NMR) for the query sequences. The drawback is that with multiple genomic initiatives currently under way the amount of raw sequence data being generated outnumbers the number of protein structures being solved each year. As of 2010 the number of structures available in PDB was over 66,000 whereas the number of protein sequences in NCBI exceeded 30,000,000 [8]. So in reality, for most sequences that are
generated today, comparing solved structures is not an option. Hence there is need for robust models built on sequence data that help us understand the evolution of proteins.

Computational methods directed at understanding evolution of proteins fall broadly into two categories. The first is phylogenetic analysis of sequences, and the second is based on prediction of structure from sequence. Phylogenetic analysis is based on robust statistical methods that identify positions in a sequence that are important in the evolution of the corresponding protein without using any information about the structure of the protein [9,10]. From a phylogenetic analysis one can compute the selection pressure on a sequence by estimating the ratio of non synonymous (dn) to synonymous substitution (ds) in the sequence at any given position. Sites that provide an adaptive advantage to the protein come under positive selection and are marked by a dn/ds ratio > 1 [11]. This approach helps us identify sites important to adaptation in evolutionary related proteins. The main limitation lies in interpreting the values of dn and ds, as these values are shown to change at different rates with time [12]. Also, single base substitutions do not tell us much about structural changes, which in most cases arise due to modifications over a region of the protein, rather than at a single site.

Prediction based computational methods predict proteins’ tertiary structures and then compare them to identify structural changes. Most commonly tertiary structures are modeled based on homologous structures in PDB [13,14,15,16,17]. The underlying concept of comparative modeling is that proteins that are evolutionary homologous usually share similar conformations and hence experimentally solved crystal structures in PDB can be used as a starting point to model structures of a homologous protein
sequence [14]. Amino acids in an alignment of homologous protein are expected to occupy similar positions in a solved tertiary space [18]. Multiple methods are available online at different web based servers that use the query sequence to locate a close match in the PDB database based on homology between sequences. It is believed that a minimum match of 30 – 40 % sequence identity is enough to successfully model a protein based on homology, as these proteins would most likely share common structural folds [18]. This belief led to the Paracelsus Challenge in 1997 in which a prize was offered for converting one protein fold to another while retaining at least 50% sequence identity [19]. This challenge has been met at least four times with the latest attempt from Alexander et al (2007) producing a pair of sequences with 88% sequence identity with completely different folds (3-alpha and alpha/beta) [20]. These results show that while homology based modeling does work, one needs to be wary not only at levels of low sequence identity. The other drawback of homology based modeling is that with multiple genomic projects coming to a conclusion, the closest homology matches sometime, are other undefined proteins (proteins without known structure) in the sequence database; hence the method cannot be applied.

A different computational method of tertiary structure prediction is called ab-initio modeling [21]. Methods that carry out ab-initio folding do so by minimizing free energy of the protein while exploring all possible folds for the sequence. These methods are very computationally intensive and can only be applied successfully to small protein sequences of <100 – 150 amino acids in length [21]. These different types of tertiary modeling are evaluated in the CASP (Critical Assessment of techniques for protein
Structure Prediction) event which is run every two years [22]. CASP started in 1994; its challenge involves solving tertiary structures of proteins, whose crystal structures are known but yet not published [22]. This event critically benchmarks different modeling techniques and calculates the success rate for each by comparing the tertiary models developed for the competition with the actual solved structures to assess quality. In the past few CASP competitions one of the most successful model has been ROSSETTA from David Baker’s lab, based on ab-initio folding [23]. The main drawback of this method is that it is computationally expensive to explore the entire protein fold space for large sequences. Thus methods based on ab-initio folding in their current state cannot be used to mine the large set of sequences being made available to us.

The tertiary structure of a protein is completely determined by the information available in its amino acid sequence [1]. Changes in sequence affect structure based on the type and position of the changes. Amino acid substitutions that completely change the physicochemical properties of the position are more likely to perturb the overall structure of the protein [4]. Most amino acid substitutions found in nature do not induce changes in structure and function of the protein [4,7,24]. Sequence changes at the core of the protein are more likely to affect its overall structure as compared to changes in the periphery [4]. Sequence changes are discrete (at a position) in nature whereas a corresponding change in the protein structure is continuous in nature which makes studying the effects of an amino acid substitution on a protein tertiary structure more complex [18]. Chothia and Lesk in a seminal paper demonstrated for the first time the relationship between sequence and structure [5]. They showed with multiple proteins that at > 50% sequence identity the
core still retains ~ 90% structural identity and as the sequence identity drops down to around 20% the structures start diverging in a non linear fashion [5]. This showed that there was no strict one-to-one relationship between sequence and structure. Illegard (2007) recently showed that structure was 3 – 10 times more conserved than sequence in proteins [6]. The non linear relationship between sequence and structure makes modeling of protein structure from sequence that much more difficult.

To study the evolution of the structure of proteins, one can also look at the individual secondary structural elements that make up the tertiary structures. The major secondary structural elements are the structured alpha helix and beta sheet and the loosely structured coils. The alpha helix is a right handed spiral conformation which repeats itself every 5.4Å with 3.6 amino acids per turn [18]. Beta sheets are an arrangement of two or more polypeptide chain linked by hydrogen bonds [18]. Coils are usually regions of the protein that lacks repetitive structural elements [18]. Other biophysical properties like hydrophobicity, solvent accessibility and flexibility also determine how the secondary structural elements come together to form the tertiary structure of the protein. Predicting secondary structure is computationally less intensive than predicting tertiary structures as the possible structure space is much smaller [18]. Multiple web servers are currently available that predict with an accuracy of ~70% ( Depending on query and the servers used) secondary structures of a protein from its linear sequence [25]. The three main methods of predicting secondary structural properties are Chou Fasman/GOR method, neural networks, and nearest neighbor [18]. Details of the different methods are presented in Chapter 3 of this thesis. Changes in secondary structural properties have been
implicated in adaptation of proteins - e.g. Colins et al. demonstrated that glycosidases enzyme of organisms that have adapted to higher temperatures have a decrease in values of predicted flexibility compared to organisms adapted to lower temperatures [26]. Thus it is important to identify such structural differences in evolving proteins to better address the question of protein adaptation.

Viruses are good systems to study protein structural evolution. Viruses are among the most rapidly evolving living organisms, with evolutionary rates as high $10^3 – 10^6$ /nucleotide/cycle of replication, which gives rise to high genetic diversity in viral populations and helps adapt to a new host [27,28]. It is unclear how adaptation to a new host is accompanied by changes in the structure of proteins [27,29,30]. To the best of our knowledge, change in sequence and its effect on predicted secondary structural properties have not been studied at the whole genome level between groups of phylogenetically related viruses. This leads to the question of whether changes in sequence that occur in a viral genome during adaptation to a new host are also accompanied by changes in predicted structural properties, and whether these sites can be identified using a statistical pipeline. We studied the proteins of avian and human metapneumovirus as an example of a virus recently adapted to the human host [31]. We also studied the reverse transcriptase (RT) protein in the primate lentivirus family. This protein is a functional heterodimer known to be functionally different within the lentivirus family, with HIV1 showing higher error rate of RT than other primate lentiviruses [32,33,34]. Using sequence data from metapneumovirus and lentivirus we developed a statistical pipeline to study the
structure changes of the protein at the level of secondary, tertiary, and quaternary protein structures (Chapters 3, 4, 5).

1.2 Objectives and Outline

The main objective of this thesis is to develop a multivariate statistical pipeline to identify changes in predicted secondary structure in evolving viral proteins. Predictions of structural changes in groups of homologous proteins have been performed before, but this is the first study that looked at such changes using multivariate statistical tools, to the best of our knowledge. In Chapter 2 of this thesis we dealt with application of multivariate statistical tools to biological data using immunological data collected from cats singly and dually infected with Feline Immunodeficiency Virus derived from cats (FIV) and from cougars (PLV). The immune parameters measured here exhibited strong co-variation. Univariate approaches that did not account for such co-variation would potentially miss significant associations between infection state and the immunological parameters. We used Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), two commonly used multivariate statistical techniques, to separate the four groups of cats based on the immune parameters. These statistical techniques demonstrated the dynamic nature of the infection in cats, which could not be demonstrated via univariate analysis.

In chapter 3 we developed a statistical pipeline to identify amino acid sites associated with structural changes, based on predicted secondary structure data. In our analyses we used predicted values of secondary structures from a given webserver as an input. Different webservers use different underlying hypothesis and methods to predict
secondary structure from any given query sequence, and can produce different prediction. We looked at five of the most commonly used prediction servers and how their underlying assumptions and methods could affect identification of secondary structure differences between groups of related protein based on our pipeline. We further investigated how within and between group sequence variability affected the ability of our pipeline to identify predicted structural differences. We found that our pipeline was sensitive to the type of prediction server used but was robust to a wide range of within and between group variability levels.

The main question addressed in chapter 4 was whether viral proteins change predicted secondary structure elements during adaptation to a new host. To address the question we applied the statistical pipeline developed in chapter 3 to identify differences in structural properties between phylogenetically related groups of viruses. The dataset used in this study was comprised of full length genomes of avian and human metapneumovirus, a virus that has recently adapted to its human host. We used six predicted secondary structure properties to identify positions that differed structurally between the two groups (avian and human metapneumovirus). Using our pipeline we were able to identify combinations of secondary structure properties that differed between the two groups of viruses and the sites where the differences occurred. We demonstrated how these sites could be involved in protein packing based on changes in the predicted secondary structure properties. We were also able to show that changes in protein sequence did not correlate one-to-one with changes in structure.
In Chapter 5 we studied predicted structural differences in the reverse transcriptase (RT) protein of the primate lentivirus group. We studied the primate lentivirus RT because of the availability of crystal structures, which helps put the predicted structural differences in the context of a crystal. RT also forms a functional heterodimer, allowing us to study quaternary structural properties and how they differ in phylogenetically related primate lentiviruses. The primate lentivirus family comprises the Human Immunodeficiency Viruses (HIV) and Simian Immunodeficiency Viruses (SIV). We studied two independent events of adaptation in humans (SIVcpz to HIV1 and SIVsm to HIV2) to identify changes in predicted structure in the reverse transcriptase protein. These structural differences were spread unevenly across the known domains of RT. Our analysis suggested that changes in the predicted structure in the polymerase domain differed between the two adaptation events (SIVcpz to HIV1 and SIVsm to HIV2). We also identified secondary structural properties affecting packing of the RT protein differently across the four groups studied (SIVcpz, HIV1, SIVsm, HIV2). Analysis based on quaternary structures could separate the two adaptation events studied, but differences could not be made within the lineages. This work is currently at a preliminary stage with more efforts needed to demonstrate how the identified secondary structure changes can be informative to tertiary and quaternary structure changes.

Overall the statistical pipeline developed in this thesis can be applied to any two groups of phylogenetically related proteins to identify predicted structural changes. This pipeline, when used in conjunction with phylogenetic analysis, can help us understand
adaptive structural changes taking place over the course of evolution in different proteins from readily available sequence data.
Chapter 2

Multivariate Statistical Analyses Demonstrate Unique Host Immune Responses to Single and Dual Lentiviral Infection

2.1 Introduction

Infections with both human and feline immunodeficiency viruses cause progressive deterioration of immune responses and are characterized by decline of CD4 cells. Although circulating CD4 T cell count is an excellent indicator of disease progression, the mechanisms involved with control of HIV and FIV and maintenance of immunological competency are still under active investigation. In part, this is because host and virus factors that affect disease outcome are complex and change over time, which makes it difficult to identify specific parameters responsible for immunological health in infected individuals. However, there are several lines of evidence that suggest protection from the pathogenic consequences of HIV-1, SIV or FIV is achievable.

First, there are intriguing reports that the asymptomatic period of HIV-1 infected persons may be prolonged in individuals concurrently infected with distantly related HIV-2 [35]. Second, the most effective lentivirus vaccine trials to date are with modified live viruses in the rhesus macaque model [36]. Last, primary infection of cats with FIV derived from naturally infected wild cougars also confers protection against disease caused by virulent domestic cat FIV [37,38]. Significantly, in the SIV/macaque experimental model and the natural FIV cat system, protection can occur without clear indicators of, or correlation with, specific anti-viral responses [37,39,40]. Although
preventing infection is a principal goal, understanding how a primary infection with an attenuated or genetically distant virus can ameliorate the pathological consequences of a more virulent virus would have significant impact on therapeutic strategies. FIV infection of cats provides an important animal model to address this significant question because cats are the only hosts that develop an immunodeficiency syndrome from a natural lentivirus infection that can be used for experimental studies.

Previous studies have established that a primary infection with cougar-derived strain PLV-1695 (PLV hereafter) protects cats from CD4 decline caused by subsequent infection with a virulent feline immunodeficiency virus (FIVfca) strain (FIVC36; FIVC hereafter) [37]. Both experimental and natural infections of cats with FIVC lead to a fatal immunodeficiency syndrome [41] similar to that observed in humans infected with HIV-1 [42,43]. PLV infected cats develop low level viremia after an initial peak of virus replication [44,45] but there are no clinical manifestations of PLV infection in domestic cats or in the natural cougar host [46,47]. Though primary PLV infection did not elicit antibody or cellular adaptive responses that provided protection to FIVC infection, of importance, CD4 cell depletion was abrogated in dual, but not single, FIVC infection [37,45]. Initial analyses of a complex data set comprised of relevant immune soluble and cellular factors indicated that IFNγ differed significantly among cats with single and dual infections at some time points. However, no other clear association of immune parameters with viral infection status was detectable using univariate statistical methods [37].
We undertook the present study because, despite the broad recognition of the dynamic and inherently multivariate nature of the immune response, immunological studies often resort to simple univariate statistical techniques (such as t-tests or analysis of variance) applied separately to each response parameter in the experimental data [48]. Such analysis may obscure, or fail to reveal, important features of the data. A rich set of multivariate analysis methods exists in the statistical literature and they are frequently applied in other fields (e.g. ecology or environmental sciences) to investigate complex systems (e.g. comprising individuals or molecules with overlapping functions, amplifying downstream effects of small perturbations, etc.). Indeed, greater insight into the immune response is reached where appropriate statistical analyses are employed [49,50,51,52].

It is of great therapeutic importance to understand the immune environment that is associated with disease attenuation in chronic lentivirus infections. Thus, we conducted a multivariate analysis on the measured immune parameters collected during the first four weeks following inoculation with virulent FIVC into naive and PLV infected animals. The clinical parameters of this study, outlined briefly here, are reported in detail in Terwee et al. [37]. Our aim in the present analysis was to ascertain whether the primary immune response to FIVC was affected by the presence of PLV using a suite of multivariate statistical methods that are suitable to explore data for typical experimental infection studies, those with small sample size, missing values, and co-variation among measured parameters.

We pursued this with a three-fold strategy: (i) we quantified and characterized the co-variation among responses, (ii) we investigated significant effects of infection and co-
infection on these responses, and, (iii) we identified combinations of responses that discriminate among infection groups. By drawing composite immunological profiles of the different infection groups, our analyses demonstrate that these profiles do indeed distinguish groups at several time points during the first weeks of infection with virulent FIVC. Importantly, we identify an immunological response in dual infected cats that is unique from those mounted to either single infections with exotic PLV or virulent FIVC that suggests underlying mechanisms for disease abrogation.

2.2 Materials and Methods

2.2.1 Study Design and data Pre-processing

The experimental design is described in detail by Terwee et al. [37]. Two groups of 10 cats were sham inoculated or inoculated with PLV at day 0. At day 28, five cats in each group were inoculated intravenously with FIVC. Thus, we investigated four infection groups, each containing five animals: two single infected groups (one with FIVC and one with PLV), one group infected with both PLV and FIVC (dual infected) and one uninfected group. Using the approaches described in [37], eleven immune response parameters were measured on the 20 cats; namely, CD25, neutrophils, CD8, CD4, TNFα, Lymphocytes, IL12, IL10, IL4, IFNγ and FAS (see Table 2-1 for a list and descriptions). This was performed repeatedly at four time points post-FIVC infection; namely, days 31, 37, 52 and 59, referenced to the start of the experiment; these are days 3, 9, 24 and 31 post-FIVC challenge. Values for the immune parameters are also available at day 45, but we excluded this day because 5 of the 20 cats were missing cytokine measures.
Table 2-1. List of immunological response parameters used in the analysis.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Th-1 and Th-2 cytokines:</td>
<td>Soluble factors modulating innate and adaptive immune response</td>
</tr>
<tr>
<td>IL-4</td>
<td>B-cell growth factor, ‘Th2’ cytokine</td>
</tr>
<tr>
<td>IL-10</td>
<td>B-cell survival and proliferation, ‘Th2’; Generally antagonistic to TNF-α</td>
</tr>
<tr>
<td>IL-12</td>
<td>Stimulates production of IFN-γ and TNF-α, ‘Th1’</td>
</tr>
<tr>
<td>TNF-α</td>
<td>Stimulates systemic inflammation, regulates apoptosis, neutrophil chemotactant</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>Proinflammatory cytokine, stimulates IL-12 and TNF-α, antagonistic to IL-4, ‘Th1’</td>
</tr>
<tr>
<td>FAS</td>
<td>‘Death receptor’, induces apoptosis</td>
</tr>
<tr>
<td>Circulating immunocytes:</td>
<td>Peripheral markers of immune homeostasis</td>
</tr>
<tr>
<td>Lymph</td>
<td>T and B Lymphocytes, NK cells and monocytes</td>
</tr>
<tr>
<td>CD4</td>
<td>Cell surface marker for T helper cells (lymph subset)</td>
</tr>
<tr>
<td>CD8</td>
<td>Cell surface marker for cytotoxic T cells (lymph subset)</td>
</tr>
<tr>
<td>CD25</td>
<td>Cell surface marker for activated T cells (both CD4 and CD8) and T regulatory cells</td>
</tr>
<tr>
<td>Neutrophils</td>
<td>Neutrophil granular leukocytes, phagocytic (innate immune system)</td>
</tr>
</tbody>
</table>

This dataset was not complete for each value at each time point due to limitations of sample availability. From a total of 55 values for each infection group at each time point; day 31 had 14 (of 55) missing values for the uninfected group and 7 (of 55) missing from the FIVC group; day 37 had no missing values; day 52 had 1 (of 55) missing value for the FIVC group and 1 (of 55) for the PLV group; day 59 had 6 (of 55) missing values for the PLV group. We used Nonlinear Iterative Partial Least Squares (NIPALS; see http://biomserv.univlyon1.fr/dray/files/softwares/nipals.R [53] to impute missing values) – the nipalsPCA function from the pcaMethods package of the R language [54] produced a full ‘reconstructed’ data set.

Since our immune responses are expressed in different measurement units and present very different spreads, we also applied a data Normalization at the outset; for each day and each response, we subtracted the mean and divided by the standard deviation. This was done for all 20 cats, and thus represents an overall re-centering and re-scaling of the data, which does not affect relative positions and spreads of the four infection groups.
2.2.2 Statistical Analyses

To address challenges set out in the Introduction, we employed a few broadly used multivariate techniques, which we summarize here and detail below: (i) The levels of many pro-inflammatory cytokines and immunocytes in our data do co-vary, i.e. have a systematic tendency to move upward or downward of their averages together (positive co-variation) or in opposite directions (negative co-variation). The nature of this co-variation is informative, and can be quantified and characterized through Pearson’s correlation coefficients. In addition to studying the pair-wise correlations, we also performed a spectral decomposition of the Pearson’s coefficients matrix – which is equivalent to applying Principal Component Analysis (PCA [55,56]) to the normalized data. (ii) To investigate significant effects of single and dual infection on the responses, we used Multivariate Analysis of Variance (MANOVA; [55,56]) with tests based on the Pillai’s trace [57]. Since the responses have a marked co-variation structure, these provide enhanced power relative to univariate tests assessing differences in infection group means separately for each response [58]. (iii) To identify combinations of responses that discriminate among infection groups, we used Linear Discriminant Analysis (LDA; [55,56]. Instead of analyzing the responses as a function of the infection groups, this analyzes the groups, as labeled, as a function of the responses – which here take the role of a vector of discriminating features. LDA produces linear combinations of immune response parameters, interpretable as “immune profiles”, which best distinguish among infection status.
2.2.3 Principal Component Analysis

PCA is a method to extract directions, i.e. linear combinations of variables, which are most relevant to the variability of a multidimensional data cloud. PCA is based on the spectral decomposition of the variance-covariance matrix of the data and produces a set of orthogonal eigenvectors, each identifying a direction, associated with eigenvalues listed in decreasing order. The eigenvalues represent variances of the data along the directions described by the corresponding eigenvectors/linear combinations. Thus, the first eigenvector (linear combination) identifies the direction of largest variability, the second the one of next largest variability, subject to the constraint of being orthogonal to the first, etc (see Figure 2-1). In addition to the pair-wise measures of linear association provided by the entries of the variance-covariance matrix, the relative size of its leading (largest) eigenvalues capture the degree of linear interdependence among the variables as a group. For the analysis of immunological responses, interdependencies can be crucial since the immune system has many feedback loops that cause cell and cytokine levels to be associated.

PCA is used as a dimension reduction and data visualization technique. If a few leading eigenvalues strongly dominate all others, the data is well captured (in terms of its variability structure) by a projection on the space spanned by the corresponding eigenvectors. Consequently, in many applications, PCA can be used to produce 2 or 3-dimensional plots that provide a satisfactory visualization of high-dimensional data. PCA for each of four days in our study was carried out on the normalized data (i.e. in terms of
the correlation matrix) using the princomp function from the stats package of the R language [54].

### 2.2.4 Multivariate Analysis of Variance

MANOVA is the multivariate analogue of an Analysis of Variance (ANOVA) model, and allows a simultaneous comparison of the means of several responses across “treatment groups”, taking into account their co-variation – as represented by the error variance-covariance matrix in the MANOVA model. Tests for mean differences between two groups are usually based on Hotelling’s $T^2$ statistic. When the problem comprises more than two groups, as is the case in our study, several statistics can be used, including Pillai’s trace, Wilk’s $\lambda$, Lawley-Hotelling’s trace, Roy’s Greatest root, etc [58]. The resulting tests are generally in accord with each other; we focused on Pillai’s trace, since it provides multivariate tests that are robust to differences among the error variance-covariance matrices within groups, as well as to departures from the assumption of multivariate normality [57,58].

When assessing group differences in a multivariate setting, MANOVA can avoid misleading conclusions associated with the use of univariate t- or F-tests for individual responses. Intuitively, the reasons for this are that (a) univariate tests do not account for co-variation among responses, and (b) in addition to better power due to accounting for co-variation, multivariate tests allow a better handle of false positives (whose joint probability increases when performing multiple univariate tests) [54,59].

Two-way MANOVA models (PLV infection = Yes, No, crossed with FIVC infection = Yes, No, including overall level, main effects of each infection and interaction
effect) were fitted for each of the four days on the normalized data using the \texttt{manova} function from the \texttt{stats} package of the R language \cite{54}. In particular, for each day and each effect, we computed Pillai’s trace-based p-values. For comparison with the MANOVA, we also fitted two-way ANOVA models separately for every immunological response using the \texttt{aov} function in R. In particular, for each day, response and effect, we computed F-based p-values. This produced 11 ANOVA p-values (one for every response) for each MANOVA p-value, so the former were adjusted with a simple Bonferroni correction (i.e. multiplied by 11).

2.2.5 Linear Discriminant Analysis

Like PCA, LDA is a method to extract directions, i.e. linear combinations of variables. However, in LDA the directions are selected based on their ability to separate labeled groups in a multidimensional data cloud \cite{60}. The spectral decomposition of a matrix is employed also here; while PCA uses the variance-covariance matrix of the data, LDA uses the between groups variance-covariance matrix normalized against the within groups variance-covariance matrix. Again, the technique produces a set of orthogonal eigenvectors (each identifying a direction) associated with eigenvalues listed in decreasing order; the first eigenvector (linear combination) identifies the direction of maximal group separation, the second the one of next maximal separation, subject to the constraint of being orthogonal to the first, etc., where group separation is defined benchmarking between group against within group variation.

Whereas with MANOVA we model our immunological parameters as a function of the infection status, with LDA we analyze infection status as a function of the
immunological parameters, which here represent discriminating features. LDA, therefore, allows us to extract ‘‘immune profiles’’ (combinations of the response parameters) that provide maximal separation among infection status. LDA, like PCA, can be used for dimension reduction and data visualization. If a few leading eigenvalues strongly dominate all others, the data is well captured (in terms of group separation) by a projection on the space spanned by the corresponding eigenvectors.

LDA for each of the four days was carried out on the normalized data using the lda function from the MASS package of the R language [60]. Since our focus was on the association between immune response parameters and infection status, we produced projective graphics called bi-plots obtained through LDA. These LDA bi-plots (see Figure 2-2) comprise: (i) Projections of the data points, marked by groups, on the plane spanned by the first and second LDA directions. (ii) Elliptical contours capturing each group’s variability in the plane of the bi-plot. (iii) Loadings for the 11 immune response parameters, i.e. coefficients with which the responses enter LDA directions (which help us interpret the ‘‘immune profiles’’ we extracted); each response is represented as an arrow in the bi-plot, with orientation expressing signs, and horizontal and vertical sizes proportional to the loadings of the feature relative to the first and second LDA directions, respectively. (iv) Eigenvalues associated with LDA directions, i.e. the discriminatory ‘‘contribution’’ of these directions; each eigenvalue is represented as a bar in an inset graph – black bars are eigenvalues for first and second directions, which span the plane of the bi-plot, and the white bar is the eigenvalue for the third direction, which is not rendered in the bi-plot (since there are four groups in our study, all subsequent
eigenvalues are 0 by construction). In addition to standard plotting commands, the functions s.class and s.arrow from the ade4 package of the R language [60] were used in implementing the bi-plots.

2.3 Results

2.3.1 Immune response parameters co-vary

We expect substantial interdependencies for the immunological parameters in our study (Table 2-1). For instance, some responses will positively co-vary because they are subsets (CD4 and CD8 cells are both lymphocytes), or if there is a primary B cell proliferative response – Th2 – (IL4 and IL10), or a pro-inflammatory response – Th1 – (IL12, TNFα, IFNγ, FAS). Moreover, cytokines classified as ‘Th1’ often will negatively co-vary with those classified as ‘Th2’, but in some cases cytokines such as IFNγ and IL10 will positively co-vary [61] due to co-expression associated with regulation of the immune response. Thus, individual cytokine levels can be less informative on the nature of the immunological response at any time point than the suite of cells and cytokines.

The 11 immune parameters we considered do indeed present strong co-variation at each of the four times. We summarized pairwise linear associations computing the four correlation matrices (Pearson’s correlation coefficients), and performed PCA on each. Figure 2-1, upper panel, shows box-plots of the absolute values of the correlations at days 31, 37, 52, and 59 – the mean (median) correlations at these days are 0.39 (0.34), 0.33 (0.27), 0.27 (0.17) and 0.33 (0.31), respectively. The absolute correlations are greater than 0 with numerous large values; the majority of these correlations are statistically significant (see Tables 2-2 and 2-3). Figure 2-1, lower panel, shows the shares of
variability explained by each principal component at days 31, 37, 52, and 59. The leading eigenvalues dominate strongly in each day, with percent of explained variability for first (second) components equal to 47% (17%), 40% (24%), 36% (19%) and 36% (26%), respectively. This indicates a large degree of linear interdependence among the variables as a group, at all four times. The bulk of the (absolute) correlation values and the dominance of the leading eigenvalues decrease from day 31 over to day 52. In particular, days 52 and 59 present the lowest level of co-variation, while having a few very strong pair-wise associations.
Figure 2-1. (A) Box-plots of (absolute) pair-wise correlations and (B) Scree plots from PCA. (A) The horizontal axis indicates the four days of analysis, and the vertical axis shows the distribution of absolute pair-wise correlations among the 11 immunological response parameters. Open circles indicate parameter pairs with particularly strong correlations ("outliers" in the box-plots). Note that on day 52, the majority of parameters have lower correlations overall, but there are several that are highly correlated (anti-correlated). (B) The ordered PCA components are shown on the horizontal axis (EV stands for eigenvector). The vertical axis shows the share of overall variability explained by each component, in different colors for each of the four days of analysis.

We further explored the ten strongest pair-wise correlations at each time to identify immunological parameters with the most marked associations (Table 2-2; this represents approximately the top 20% of the correlations in each day – the reported values are all statistically significant, except for two, as noted in Table 2-2). There is
strong positive correlation between total lymphocyte counts and both CD4 and CD8 counts at all time points. CD4 and CD8 levels also show strong positive correlation at all time points.

We note that CD4 and CD8 are cell surface receptors expressed on T lymphocytes, and that CD25 can be expressed on both CD4 and CD8 cells if they are activated. Thus, strong positive correlations of these cellular markers are to be expected if there is a general expansion of T lymphocyte subsets. In fact, the strongest correlations occur among cellular parameters for days 31, 37, and 52. The correlation between FAS and IFNγ is consistent and increases with time from the FIVC infection, peaking at day 59 (bold in Table 2-2).

Table 2-2. Pearson’s coefficients of the ten strongest (absolute) pair-wise correlations between immunological responses for each day of analysis.

<table>
<thead>
<tr>
<th>Day 31</th>
<th>Correlation Pair</th>
<th>Day 37</th>
<th>Correlation Pair</th>
<th>Day 52</th>
<th>Correlation Pair</th>
<th>Day 59</th>
<th>Correlation Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.603</td>
<td>FAS, IFNγ</td>
<td>0.574</td>
<td>IL4, IL12</td>
<td>0.315</td>
<td>Lymph, TNFα</td>
<td>0.469</td>
<td>IFNγ, CD25</td>
</tr>
<tr>
<td>0.605</td>
<td>IFNγ, CD8</td>
<td>0.638</td>
<td>IL10, IL12</td>
<td>0.355</td>
<td>IL10, Lymph</td>
<td>0.486</td>
<td>IL10, Lymph</td>
</tr>
<tr>
<td>0.702</td>
<td>IFNγ, TNFα</td>
<td>0.718</td>
<td>IL12, TNFα</td>
<td>0.480</td>
<td>IL10, TNFα</td>
<td>0.567</td>
<td>IFNγ, IL4</td>
</tr>
<tr>
<td>0.808</td>
<td>CD8, CD12</td>
<td>0.718</td>
<td>IL10, TNFα</td>
<td>0.722</td>
<td>FAS, IFNγ</td>
<td>0.583</td>
<td>FAS, Neutr</td>
</tr>
<tr>
<td>0.820</td>
<td>CD4, CD25</td>
<td>0.754</td>
<td>CD4, CD25</td>
<td>0.829</td>
<td>Lymph, CD25</td>
<td>0.583</td>
<td>FAS, TNFα</td>
</tr>
<tr>
<td>0.831</td>
<td>Lymph, CD25</td>
<td>0.792</td>
<td>CD8, CD25</td>
<td>0.848</td>
<td>CD8, CD25</td>
<td>0.645</td>
<td>IFNγ, Neutr</td>
</tr>
<tr>
<td>0.836</td>
<td>IL10, TNFα</td>
<td>0.828</td>
<td>Lymph, CD25</td>
<td>0.858</td>
<td>CD4, CD25</td>
<td>0.777</td>
<td>CD4, CD8</td>
</tr>
<tr>
<td>0.847</td>
<td>Lymph, CD4</td>
<td>0.881</td>
<td>Lymph, CD8</td>
<td>0.888</td>
<td>Lymph, CD8</td>
<td>0.778</td>
<td>Lymph, CD8</td>
</tr>
<tr>
<td>0.873</td>
<td>Lymph, CD8</td>
<td>0.910</td>
<td>Lymph, CD4</td>
<td>0.941</td>
<td>Lymph, CD4</td>
<td>0.836</td>
<td>FAS, IFNγ</td>
</tr>
<tr>
<td>0.943</td>
<td>CD4, CD8</td>
<td>0.966</td>
<td>CD4, CD8</td>
<td>0.941</td>
<td>CD4, CD4</td>
<td>0.898</td>
<td>Lymph, CD4</td>
</tr>
</tbody>
</table>

*Pair-wise correlations that increase (bold) or decrease (italics) are highlighted.

All coefficients are statistically significant (p-values substantially smaller than 0.05) using a test for association/correlation between paired sample, except for the two negative correlations of Lymph with TNFα and IL10 at day 52 (TNFα p-value = 0.13, IL10 p-value = 0.12).

Since our data comprises only 20 observations (cats), we also considered Spearman’s correlation coefficients, which are a robust version of Pearson’s, computed on the ranks instead of the measured levels. Similar to the Pearson’s, the mean (median)
Spearman’s correlations at days 31, 37, 52 and 59 are 0.41 (0.39), 0.29 (0.21), 0.27 (0.24) and 0.32 (0.32), respectively. Exploring the ten strongest Spearman’s pair-wise correlations at each time (Table 2-3) also produces results consistent with Pearson’s (Table 2-2).

It is important to note that our co-variation analyses concern the complete dataset from all 20 cats and do not capture or exploit specific relationships of immunological parameters with the infection groups; therefore the strong interdependencies detected through both pair-wise correlations and PCA confirm the need for multivariate techniques in investigating relationships between these immune parameters and infection status.

**Table 2-3.** Spearman’s coefficients of the ten strongest (absolute) pair-wise correlations between immunological responses for each day of analysis.

<table>
<thead>
<tr>
<th>Day 31 Correlation Pair</th>
<th>Day 37 Correlation Pair</th>
<th>Day 52 Correlation Pair</th>
<th>Day 59 Correlation Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.67 CD4, CD25</td>
<td>0.51 IL10, TNFα</td>
<td>0.50 IL10, CD4</td>
<td>0.51 CD4, Neut</td>
</tr>
<tr>
<td>0.56 IL10, TNFα</td>
<td>0.53 IL10, IL12</td>
<td>0.54 Lymph, CD25</td>
<td>0.51 FAS, CD25</td>
</tr>
<tr>
<td>0.66 IFNγ, CD8</td>
<td>0.65 FAS, IFNγ</td>
<td>0.54 Lymph, CD8</td>
<td>0.52 Lymph, Neut</td>
</tr>
<tr>
<td>0.61 Lymph, CD25</td>
<td>0.88 CD8, CD25</td>
<td>0.61 FAS, IFNγ</td>
<td>0.60 FAS, Neut</td>
</tr>
<tr>
<td>0.78 CD8, CD25</td>
<td>0.89 Lymph, CD4</td>
<td>0.64 CD4, CD25</td>
<td>0.61 CD4, CD8</td>
</tr>
<tr>
<td>0.79 Lymph, CD8</td>
<td>0.90 CD4, CD25</td>
<td>0.68 CD8, CD25</td>
<td>0.66 IFNγ, Neut</td>
</tr>
<tr>
<td>0.80 IFNγ, TNFα</td>
<td>0.92 Lymph, CD25</td>
<td>0.72 Lymph, CD4</td>
<td>0.84 Lymph, CD4</td>
</tr>
<tr>
<td>0.89 CD4, CD8</td>
<td>0.97 CD4, CD8</td>
<td>0.79 CD4, CD8</td>
<td>0.88 FAS, IFNγ</td>
</tr>
</tbody>
</table>

*Pair-wise correlations that increase are bolded. The sign indicates whether a pair co-varies positively or negatively. All coefficients are statistically significant (p-values substantially smaller than 0.05) using a test for association/correlation between paired sample.

2.3.2 Mean immune responses differ significantly across infection status

To investigate significant effects of infection and co-infection on the immune response parameters and more specifically to assess differences between group means, we fit a MANOVA model for each of the four days considered. The model expresses our
vector of 11 responses as a function of the PLV and FIVC infection status in a 2-way scheme comprising the four infection groups. In this manner, the effect of the two PLV-infected groups (PLV infection only and dual infection) can be compared with non-PLV infected groups (FIVC infection only, and uninfected) and vice versa. Tests based on Pillai’s trace statistics can therefore be performed for the simple effect of PLV, the simple effect of FIVC and the “interaction” effect. Table 2-4 contains estimated effects and corresponding p-values.

Table 2-4. Estimated effects (E) and Pillai’s trace p-values (p) for a two-way MANOVA scheme comprising simple and interaction effects of PLV and FIVC, for each day of analysis.

<table>
<thead>
<tr>
<th></th>
<th>Day 31</th>
<th>Day 37</th>
<th>Day 52</th>
<th>Day 59</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLV</td>
<td>p = 0.006*</td>
<td>p = 0.16</td>
<td>p = 0.01</td>
<td>p = 0.18</td>
</tr>
<tr>
<td></td>
<td>E = 0.95</td>
<td>E = 0.81</td>
<td>E = 0.93</td>
<td>E = 0.80</td>
</tr>
<tr>
<td>FIVC</td>
<td>p = 0.52</td>
<td>p = 0.07</td>
<td>p = 0.02</td>
<td>p = 0.04*</td>
</tr>
<tr>
<td></td>
<td>E = 0.65</td>
<td>E = 0.86</td>
<td>E = 0.91</td>
<td>E = 0.89</td>
</tr>
<tr>
<td>PLV × FIVC</td>
<td>p = 0.96</td>
<td>p = 0.29</td>
<td>p = 0.18</td>
<td>p = 0.56</td>
</tr>
<tr>
<td></td>
<td>E = 0.36</td>
<td>E = 0.75</td>
<td>E = 0.80</td>
<td>E = 0.63</td>
</tr>
</tbody>
</table>

*Significant results are in bold.
**This value is not significant if the MANOVA is run without imputing the values.

The simple effect of FIVC is significant at day 52 and 59 but not at days 31 and 37. Conversely, the simple effect of PLV has the highest significance at day 31, is significant at day 52 and is not significant at days 37 and 59. Our MANOVA testing does not detect a significant interaction effect at any of the times considered. However, the interaction p-value is smallest at day 52. We also conducted the MANOVA on the data without missing value imputation. This decreased the sample size by different amounts for each day (See Methods). The p-value for the effect of FIVC at day 59 increased slightly (from 0.04 to p= 0.06), and the p-value for the interaction effect of PLV and
FIVC at day 52 decreased from 0.18 to 0.06 (the omitted imputed values here involved one animal each from the FIVC and PLV group). In all, the data highlights day 52 (day 24 post FIVC infection) as the one in which multivariate responses are most affected by infection status.

For comparison with the above multivariate results, we also fit 11 univariate 2-way ANOVA models for each of the four days considered. Here the model expresses an individual response as a function of the PLV and FIVC infection factors in the same 2-way scheme, and F-tests were performed for the simple effect of PLV, the simple effect of FIVC and the interaction effect. For each effect, and each day, we therefore obtained 11 p-values – one for every response; Table 2-5 lists response names, estimated effects and p-values for the cases that remained significant after Bonferroni correction.

Table 2-5. Estimated effects (E) and Bonferroni corrected F-based p-values (p) for two-way ANOVA schemes applied to each immunological parameter separately, for each day of analysis.

<table>
<thead>
<tr>
<th></th>
<th>Day 31</th>
<th>Day 37</th>
<th>Day 52</th>
<th>Day 59</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLV</td>
<td>IFNγ&lt;sup&gt;*&lt;/sup&gt;</td>
<td>IFNγ&lt;sup&gt;+&lt;/sup&gt;</td>
<td>IFNγ&lt;sup&gt;+&lt;/sup&gt;</td>
<td>IFNγ&lt;sup&gt;+&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>p = 0.026; E = 0.45</td>
<td>p = 0.001; E = 0.62</td>
<td>p = 0.05; E = 0.40</td>
<td>p = 0.02; E = 0.48</td>
</tr>
<tr>
<td>FIVC</td>
<td>None</td>
<td>CD25&lt;sup&gt;+&lt;/sup&gt;</td>
<td>IFNγ&lt;sup&gt;+&lt;/sup&gt;</td>
<td>CD25&lt;sup&gt;+&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>p = 0.0002; E = 0.70</td>
<td>p = 0.02; E = 0.47</td>
<td>p = 0.05; E = 0.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD8&lt;sup&gt;-&lt;/sup&gt;</td>
<td>FAS&lt;sup&gt;-&lt;/sup&gt;</td>
<td>Neutrophils</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p = 0.00005; E = 0.74</td>
<td>p = 0.02; E = 0.46</td>
<td>p = 0.0002; E = 0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD4&lt;sup&gt;-&lt;/sup&gt;</td>
<td>IFNγ</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p = 0.0005; E = 0.66</td>
<td></td>
<td>p = 0.001; E = 0.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lymphocytes</td>
<td>FAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p = 0.0005; E = 0.66</td>
<td></td>
<td>p = 0.03; E = 0.48</td>
<td></td>
</tr>
<tr>
<td>PLV × FIVC</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

Only significant effects are reported, with the name of the corresponding parameters.

This value is not significant if ANOVA is run without imputing the values (p < 0.13).

Similar to the MANOVA results, interaction effects are not detected as significant for any of the responses, at any of the times considered. However, a number of simple
effects of PLV and FIVC status do survive the correction for multiple testing at each time. In particular, the effect of PLV is always significant in terms of the IFNγ levels although significance declines at days 52 and 59 in this group. In contrast, the effect of FIVC is significant in terms of both IFNγ and FAS at days 52 and 59 (recall these responses present a very strong positive correlation at these times; Table 2-1).

In addition, cell phenotype parameters indicative of activated lymphocytes (CD25, CD8, CD4, and lymphocyte) at day 37, and neutrophil counts at day 59, show a significant effect of FIVC. Thus, the ANOVA tests detect the strongest effect of FIVC infection at days 37 and 59 and indicate IFNγ as a “driver” of the immune response to PLV infection at all time points. The role of the immunophenotype is confined to day 37 (day 9 post-FIVC infection) as a simple effect of FIVC infection. Also here, we conducted the ANOVAs without missing value imputation, obtaining similar results, with the exception PLV did not have a significant effect for IL4 when omitting the missing values. In conclusion, both ANOVA and MANOVA tests suggest that there are differences in immune responses of groups with primary FIVC and PLV infection. However, some critical differences exist in the outcome of univariate and multivariate analysis approaches, which will be interpreted further in the discussion section.

2.3.3 Specific immune profiles, which evolve over time, discriminate among infection status

To identify combinations of responses that discriminate among the infection groups (“immune profiles”), we performed LDA for the four groups on the 11 responses, for each of the four days considered. LDA output for days 31, 37, 52, and 59 is
summarized in the four bi-plots in Figure 2-2. On day 31 (3 days post-FIVC infection; upper left panel) there is essentially one discriminating signal in the data, with a strongly dominating eigenvalue associated with the first LDA axis, which separates PLV infected (single and dual) and non-PLV infected (uninfected and FIVC infected) cats. The feature loadings indicate that this corresponds to a ‘‘trade-off’’ between CD8 and CD4 cell numbers, with other immunological parameters playing secondary roles. It is also noteworthy that PLV infection induces an elevation in neutrophil count, CD25 cells, IFNγ, and IL12 compared to uninfected or recently FIVC-infected cats at this time point. Note that while some separation between FIVC-infected and uninfected cats, and single and dual PLV infected cats, exists along the second LDA axis, the eigenvalue associated with this axis is much smaller than the first.
Figure 2-2. Bi-plots from LDA, for days 31 (upper left), 37 (upper right), 52 (lower left) and 59 (lower right). The horizontal and vertical axes on each panel represent first and second LDA direction, respectively. The arrows depict loadings of the 11 responses; their orientation and length indicate the role of each immunological parameter in relation to the LDA directions and the different infection groups, on each of the analysis days. The insets in each panel show first, second and third LDA eigenvalue (discriminatory importance of the corresponding LDA directions).

By day 37 (9 days post-FIVC infection; Fig. 2-2, upper right panel), the picture has become more complex. The data contain at least two discriminating signals (see eigenvalues in the in-set), and the four groups separate along both the first and the second LDA axes. There is complete separation of single PLV- and FIVC infected cats along the first LDA axis due primarily to higher expression of CD8, CD25 and FAS in PLV
infected cats. FIVC dual and single infected cats show a substantial overlap in the projection on the first two LDA directions. Notably, the second LDA direction separates uninfected from all infected cats (PLV single, FIVC single and dual); uninfected cats have higher CD4, TNFα and neutrophil levels, and lower CD8, lymphocyte, IFNγ and IL10 than any of the infected groups.

The discriminating signals captured by the first and second LDA directions provide maximal resolution of the four groups at day 52 (24 days post-FIVC infection; Figure 2-2, lower left). In particular, at day 52 we observe a clear “immunological profile” for dual infected cats, as they cluster far away from uninfected and both groups of single infected cats. Based on the loadings, this profile is characterized by higher CD8 and CD25 counts, and elevated levels of IL4, FAS, and IFNγ than that of the other groups. FIVC single infections separate from uninfected cats and single PLV infections due to higher levels of IL10 and lower CD4 counts. PLV infections separate from uninfected cats principally on the basis of lower CD4 counts.

The four treatment groups are also distinguishable at day 59 (31 days post-FIVC-infection; Figure 2-2, lower right). Dual infected cats separate from the other three groups due to higher IFNγ levels and CD8 and lymphocyte counts. Single FIVC infected and dual infected cats separate along the second axis. Dual infected cats have elevated levels of TNFα, IL4, and IFNγ and CD8 cells and neutrophils compared to single FIVC infected animals. Single FIVC infections are characterized by higher IL12 and FAS levels and lymphocyte counts than dual infections. Uninfected and PLV single infected cats overlap in the projection on
the first two LDA axes; both have higher levels of CD4 and CD25 cells and neutrophils compared to cats with single and dual FIVC infection.

The ellipses in the bi-plots of Figure 2-2, which represent the variance-covariance structure of the projected data points comprising the four groups, show marked differences across groups at each time and across times for each group. In particular, in these projective representations the uninfected group often appears more variable than the infected groups. However, these ellipses are not indicative of overall variability in the 11-dimensional response space. When considering all 11 dimensions, the uninfected group is indeed the tightest, with a total variance (sum of the 11 variances) that is fairly constant over time, whereas the infected groups show higher and changing total variance (Table 2-6). This indicates that infection status may affect not just the location, but also the variability structure of the 11 immunological measures considered here.

Table 2-6. Total variances within infection status groups on different days.

<table>
<thead>
<tr>
<th></th>
<th>Day 31</th>
<th>Day 37</th>
<th>Day 52</th>
<th>Day 59</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual</td>
<td>14.26</td>
<td>12.36</td>
<td>9.29</td>
<td>10.43</td>
</tr>
<tr>
<td>PLV</td>
<td>12.54</td>
<td>7.2</td>
<td>15.95</td>
<td>9.73</td>
</tr>
<tr>
<td>FIVC</td>
<td>4.18</td>
<td>8.16</td>
<td>6.81</td>
<td>6.45</td>
</tr>
<tr>
<td>Uninfected</td>
<td>5.55</td>
<td>3.76</td>
<td>5.82</td>
<td>4.56</td>
</tr>
<tr>
<td>All cats</td>
<td>11 ²</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

²The total variance for all cats (last row) is 11 on each day because of normalization.

2.4 Discussion

In this study, we used multivariate statistical techniques to evaluate the immunological responses of cats infected with an apathogenic FIVC (FIVpco strain PLV-1695 derived from a cougar), virulent FIVC, or both viruses. FIVC infection on its own causes immunosuppression and death [41]. However, cats infected first with PLV and
then with FIVC are protected from these consequences [37]. Immunological parameters measured on the cats consisted of lymphocyte expression levels for five cytokines, total number of circulating lymphocytes and neutrophils, three functional markers displayed on T lymphocytes, and a cell surface molecule regulating cell survival. Detecting differences among infection groups, as well as the discriminatory roles of various immune parameters, is complicated by interdependencies among these parameters, their diverse kinetics, and difficulties in gauging variability given the small sample size at our disposal. Employing well established and straightforward multivariate statistical techniques such as PCA, MANOVA and LDA, despite the limited number of observations, we found clear evidence that the response variables measured in our study do distinguish the four experimental groups.

Understanding the immunological response in cats infected with both PLV and FIVC is of particular interest in this study because dual infected cats are protected from FIVC-induced disease. Our data show that simultaneous infection with the two viruses elicits an immune response that is substantially different from that mounted to uninfected or single infection with either the apathogenic PLV or the virulent FIVC. FIVC infection is initiated at day 28 post-PLV infection. Notably, the immune profile of dual infected cats resembles that of PLV, not FIVC, 3 days post-FIVC infection (day 31). As FIVC infection proceeds to day 37, the immunological profile of all but one of the dual infected cats starts to resemble, or overlap with, that of FIVC infected cats. This highlights both the variation in cat responses at this date (see Table 2-6) and the dynamic nature of the immune response to a mixed infection. By day 52 (24 days after FIVC infection) the
immunological response in dual infected cats is clearly distinct from all groups and is
dominated by an elevated level of CD8, CD25, and FAS expressing cells, and increased
numbers of lymphocytes and neutrophils. This phenotype may reflect expansion of an
important effector T cell subset in these cats.

IL4 and IFNγ are the only cytokines that contribute to the unique immune profile
of dual infected cats at day 52. The dynamic nature of host response to dual virus
infection is apparent in comparing the profiles of days 52 and day 59. For the first time,
IFNγ is clearly a driver in dual infected cats at day 59 although the overall profile is not
typically pro-inflammatory. In addition, increased lymphocyte counts with lower levels
of CD4 and average levels of CD8 cells compared to uninfected and PLV infected cats
suggest that B cells or NK cells may be increased in dual (and to a greater extent in single
FIVC) infected cats at this time.

How does the immunological profile of cats infected with FIVC in the presence of
a preexisting PLV infection differ from cats infected with FIVC alone? In general, single
FIVC infection elicits an overall increase in IL10, IL4, and IL12, but lower expression of
CD25, and CD8 throughout the first 31 days of infection (day 31– 59 of the experiment)
compared to other groups. Prior to day 59, FAS levels are lower in single FIVC infected
cats but are increased at day 59 of the experiment. Neutrophil and CD4 levels also are
lower at day 59. In contrast, the immune response in dual infected cats is an activated
cellular profile maturing over the 31 days of dual infection. It is noteworthy that the
differences in the immune response to FIVC in dual and single infected cats likely arise
during the first week of infection because after that point, the profiles of PLV and
uninfected cats are very similar. In addition, while FIVC infection induces CD4 decline and dual infected cats are protected from CD4 loss [37], CD4 is not associated with the dual infection group in the first three weeks of infection. In contrast, elevated levels of CD8 cells are consistently associated with PLV and dual infected cats over this time frame. We show that the immunological parameters considered in our study present a marked co-variation structure; an indication that multivariate statistical methods are needed to derive a clear association between infection state and immune response.

MANOVA tests evaluate whether there are significant effects of the treatment groups on the vector of response parameters as a whole. Using a two-way scheme, we demonstrate that there is a significant effect of PLV infection on immune parameters on day 31. This is also demonstrated by the LDA results from day 31, which show a clear separation of PLV and dual infected groups from FIVC and uninfected cats. Cell count responses, which have high positive correlations at all times, and in particular CD4 and CD8 counts, emerge as key differentiators between PLV and non-PLV cats at day 31. Importantly, neither CD4 nor CD8 carry significant effects in their individual ANOVAs for day 31 (Table 2-5).

Also the difference CD8-CD4, which captures the dominant trade-off shown by the LDA loadings in Figure 2-2 (upper left) has no significant effect in an individual ANOVA for this day (data not shown). Yet, in combination with the other responses, this tradeoff has a crucial role in group separation. Our MANOVA results also indicated that there were significant effects of both FIVC and PLV infection on day 52 and a comparatively weaker role of the FIVC infection at day 59, both of which are consistent
with the LDA findings. By contrast, univariate analysis (ANOVA) identifies the strongest effect of FIVC infection at days 37 and 59. Further, the immune parameters that emerge as significant in the univariate analysis (ANOVA) are not necessarily those that discriminate the four groups as determined by our multivariate approaches. For example, at day 52, when the effects of the dual infection are most pronounced based on MANOVA and LDA, ANOVA results show a marginally significant difference for IFN\(\gamma\) between PLV and non-PLV infections, and significant differences for IFN\(\gamma\) and FAS between FIVC and non-FIVC infections, suggesting that IFN\(\gamma\) is a strong driver of group discrimination. However, LDA reveals a more complex interplay among the immunological parameters at day 52. Higher levels of CD8 cell and lymphocyte and neutrophil counts, enhanced expression of FAS, IL4, and IFN\(\gamma\), and lower levels of IL10, TNF\(\alpha\) and CD4 cell counts are all associated with the separation of dual infected cats from all other groups. Most important, we are able to determine that the immunological response to single FIVC infection was relatively stable - dominated by IL-10 and low CD25 and CD8 - in contrast to the dynamic and pro-inflammatory profile seen in the dual infected cats over the first month following FIVC infection.

Although we qualitatively compared output of multivariate analyses at different time points, because of the small sample size of this study, we limited ourselves to relatively simple techniques, and did not venture into the more complex territory of multivariate methods and models that would allow us to include time as an endogenous factor in the study. We also did not attempt to account for heterogeneity we observed in the variability structure of the immune responses within each infection group. Methods
and models to deal explicitly with dynamics and heteroschedasticity do exist [62,63,64,65], and we are planning to investigate their use on more complete data sets in the future.

(This work was published in PLoS One 2009: Roy, S.; Lavine, J.; Chiaromonte, F.; Terwee, J.; VandeWoude, S.; Bjornstad, O.; Poss, M. Multivariate statistical analyses demonstrate unique host immune responses to single and dual lentiviral infection. PLoS One 2009, 4, e7359.)
Chapter 3

Developing a statistical pipeline to identify secondary structure changes between related groups of proteins

3.1 Introduction

The numbers of solved crystal structures have grown at a healthy pace over the last few decades with currently over 66,000 entries in PDB. Over the same period of time the number of protein sequences in the protein database of NCBI has exploded to well over 30,000,000 [8]. As the rate of availability of sequences outpaces the rate at which crystals are solved we need better tools to predict protein structure from sequence to study structural changes between groups of related proteins. Predicting the secondary structure of a protein from amino acid sequences goes back to the 1970s [66]. Starting from the Chou and Fasman method in the late 70s, prediction methods have improved at a rapid rate due to the increase in computing speed and in the number of solved crystal structures available. The prediction methods mostly used today fall in three broad categories; i) Chou Fasman/GOR method ii) Neural Networks iii) Nearest neighbor iv)

The Chou Fasman method analyzes the frequency of each amino acid present in alpha helices, beta sheets and coils of known structures [66]. Probability scores are determined for each amino acid based on its frequency of representation in known alpha helices, beta sheets or coils from protein structure databases. For instance, a region of a sequence rich in Alanine or Glutamine has a higher propensity to form an alpha helix. For predicting secondary structure by Chou Fasman, a nucleation site is first identified in the sequence. A nucleation site is a stretch of six or seven amino acids that show high
propensity for one of the three secondary structural properties. This nucleation site is extended along the sequence until the probability for the property drops under 0.5. The GOR (Garnier, Osguthorpe and Robson) method is a modification of the Chou Fasman method [67]. The GOR method uses a sliding window analysis which takes into account the probability of the neighboring amino acids to form a similar structure and gives a moving average as a prediction score for secondary structure [67]. This puts the amino acids in the context of the sequence and is an improvement over the Chou Fasman method.

Neural networks are one of the most computationally intensive and sophisticated methods of predicting secondary structures. Some of the most popular servers like PSIPRED and Predict Protein use neural networks to predict secondary structure [18,68,69,70]. While the Chou Fasman and the GOR methods predict structure based on a pattern of amino acids, neural networks match these same patterns of amino acids to known domains structures in the protein database. The computer algorithms based on machine learning are first trained to identify known secondary structure properties in a domain database and identify patterns of amino acids that appear in them. After learning form the training sets, the program matches an unknown sequence to its training sets and based on the matches predicts the secondary structure for the region [18,69].

Nearest neighbor methods are also a type of machine learning algorithms. Here structure prediction for an input sequence in based on sequence homology to known sequences in the database [18,71,72]. Close matches for the query sequence are first identified from the database by calculating sequence similarity based on an alignment.
From the identified matches, fragments are generated of a desired window size, which represent all possible different structures from the homologous sequences. A sliding window is then used on the query sequence to identify the best structure match from the homologous fragments for predicting structure of the query sequence. PREDATOR is a popular example of this approach [71,72]. In this study we look at commonly used servers based on these three methods (Chou Fasman/GOR, Neural Networks and Nearest Neighbor).

All the methods described above are associated with some error rate, as they are not 100% accurate. The performance of each method is very case specific and some outperform others for different query sequences. In cases of sequences where no structural information exists for similar homologous proteins, a method working on frequency of amino acid present in a window may work as well or better that one relying on various types of machine learning algorithms. On the other hand, in cases of sequences for which crystal structures exist for similar protein, methods relying on training data will most likely outperform other methods.

In our statistical pipeline we use predictions of secondary structure elements from different servers, and used them to predict the location of structural differences between phylogenetically related groups of proteins. To develop the pipeline we considered protein sequences from Alexander et al 2007 (with known crystal structures) and used them as query sequences in five different prediction servers [20]. The servers used were ExPASy (Chou Fasman/GOR), Jpred 3 (Neural Network), Predict Protein (Neural
Network), PREDATOR 2.1.2 (Nearest Neighbor), and PSIPRED (Combination of Neural Network and Nearest Neighbor) [66,67,68,69,70,71,72,73,74].

Changes in sequence affect structure, but no one-to-one relation exists. It has been shown that structure can be three to ten times more conserved than sequence [5,6]. Thus identifying structural changes within related sequences may therefore be complicated by variability in amino acid sequences, and not knowing how it translates to variability in structure. Most methods based on some sort of similarity search use any value above 45% sequence similarity as a possible match, while some push this threshold to as low as 25% [18]. In cases of sequences of recently evolved proteins we do expect sequences of high similarity. Such sequences may all be predicted to have similar structure based on homology search, and important recent adaptive structural changes may not be identified. Here, we test how sequence variability affects prediction of secondary structure from a prediction server, and in turn what effect this has on the identification of structural changes in our statistical pipeline. We used the ExPASy server to generate secondary structure prediction for groups of randomly generated sequences (based on Alexander et al 2007 sequences) with various within and between group sequence variability, and tested the effect of this variability on the identification of structural differences based on our pipeline [20,74].
3.2 Material and Methods

3.2.1 Sequence Data

We used eight sequences made available from Alexander et. al. (2007) [20]. The sequences were of the Ga (4) and Gb (4) subdomain of the protein G of Streptococcus (Fig 3-1). This protein is a multi-domain cell wall protein with domain Ga binding to human serum albumin and Gb to the Fc region of IgG. The two domains Ga and Gb, although similar in length, have two different folds, 3-α and α/β respectively (Fig 3-1). The eight sequences for this dataset are from four pairs of Ga and Gb based on increasing sequence identity. The original pairs (Ga and Gb) are 17% identical and are indicated as Ga17 and Gb17. The other three pairs have increasing identities of 30% (Ga30 and Gb30), 77% (Ga77 and Gb77) and 88% (Ga88 and Gb88) respectively. Seven amino acid positions were identified by Alexander et. al. as key to maintaining the differences in folds between Ga17 and Gb17 [20]. These sites are referred to as ‘true SInS’ (structurally informative sites) hereafter. For comparing different prediction servers, we used the original 8 sequences from Alexander et. al [20]. To test the effect of sequence variability on prediction of SInS, we generated our own sequences, as outlined later in this section.
3.2.2 Prediction of secondary structure

We predicted secondary structure properties (alpha helix, beta sheet and coils) using five widely used prediction servers (ExPASy, Jpred 3, Predict Protein, Predator 2.1.2, Psipred) [66,69,70,71,72,73,74]. The parameters used in each prediction server were as follows. In ExPASy the three structural properties were determined by the Deleage and Roux method (similar to GOR) [74,75]. The window size was set to 5, with the relative weight of the edges set to 10% of the center. The exponential weight...
variation model was used and the data normalized to range from 0 to 1. For Jpred 3 and Predict Protein (Neural Networks) no tuning parameters were specified in the input form for tuning the neural networks hence the default settings established by the authors were used for this study [70,73]. For Predator 2.1.2 (Nearest Neighbor) we used single sequence prediction from the options available [71,72]. For Psipred (Combination of Neural Network and Nearest Neighbor) we chose the Psipred ver 3.0 prediction method without the use of any filters [68,69]. As the query sequence was the same for each of the servers it was reasonable to compare the outputs.

3.2.3 Within group and between group sequence variability

The number of amino acid positions that were substituted to get the Ga88 sequence starting from the original Ga17 sequence was 24, whereas 18 substitutions were made in Gb17 to get Gb88. To generate sequence variability we performed what we call a combinatorial sequence substitution for Ga and Gb. The constraints imposed were that: (1) an amino acid position in Ga17 and Gb17 could only change to that present in Ga88 and Gb88 respectively. This allowed us to maintain the two distinct structures as best as possible; (2) the positions that were mutated were among those that actually changed between the original sequence Ga17 and Ga88 (Gb17 and Gb88). This was done to ensure that the sequence space created would not disrupt the structures of Ga and Gb and that the positions of the true SInS that were known to be different between Ga88 and Gb88 were maintained. Starting from all possible single amino acid mutations (any one position mutated out of the possible 24 positions in Ga17 and 18 possible in Gb17) we moved to all possible combinations of double mutations for Ga17 and Gb17, and then on
to higher combinations of mutation up to 24 possible mutations in Ga17 and 18 possible mutation in Gb17. From the resulting sequence space we randomly sampled 10 sequences from Ga and 10 sequences from Gb and calculated within and between group sequence variability using pairwise distance in R. This process was repeated 1000 times for both Ga and Gb. Sequence variability was computed as the number of amino acid position that differed between the two groups of proteins represented as a percentage of the total number of amino acid in the protein. Using 4 Ga sets (A1 – A4) and 3 Gb sets (B1 – B3) selected at random among the 1000 sets generated for each we sampled a within group sequence variability ranging from 9.6 – 23.2 % and between group sequence variability ranging from 28.5 – 53.7% (Table 3-1). The sequences in these 7 sets (4 Ga and 3 Gb) were used to generate six structural properties (alpha helix, beta sheet, coils, accessibility, flexibility, hydrophobicity) from the ExPASy webserver [74].
Table 3-1. Between group and within group sequence variability of 4 groups of $\text{Ga}$ (A1 – A4) and 3 groups of $\text{Gb}$ (B1 – B3) sequences. The within groups sequence variability is shown across the diagonal of this table.

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>11.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>NA</td>
<td>15.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>NA</td>
<td>NA</td>
<td>19.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>23.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>39.6%</td>
<td>53.7%</td>
<td>48.0%</td>
<td>37.5%</td>
<td>9.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>38.5%</td>
<td>52.5%</td>
<td>46.8%</td>
<td>36.3%</td>
<td>NA</td>
<td>13.5%</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>30.5%</td>
<td>44.8%</td>
<td>38.1%</td>
<td>38.1%</td>
<td>NA</td>
<td>NA</td>
<td>17.5%</td>
</tr>
</tbody>
</table>

3.2.4 Identification of structurally informative sites (SInS)

To identify structurally informative sites (SInS) we used a variation of the LDA algorithm to compute a distance metric between groups of $\text{Ga}$ and $\text{Gb}$ sequences. LDA is a method to extract the linear combinations of variables that maximally separate labeled groups in a multivariate dataset [60], using the spectral decomposition of a matrix, which represents the between groups variance-covariance normalized against the within groups variance-covariance. This produces a set of orthogonal eigenvectors (each identifying a linear combination of the original variables) associated with eigenvalues listed in decreasing order; the first eigenvector (linear combination) identifies the direction of maximal group separation, the second the one of next maximal separation, subject to the constraint of being orthogonal to the first, etc.
At each position along the protein we have 12 observations (12 metapneumovirus sequences) in 6 dimensional space (six predicted properties), divided in two groups (avian and human metapneumovirus groups). Using an LDA approach we could define the distance between two groups as amino acid position i as

\[ d_i = ((\mu_{1i} - \mu_{2i})^T \times (\Sigma_i + \alpha I)^{-1} \times (\mu_{1i} - \mu_{2i}))^{1/2} \]

where \( \mu_{1i} \) indicates the mean properties vector in group 1, \( \mu_{2i} \) the corresponding vector in group 2, and \( \Sigma_i \) the average within group covariance matrix. Note that this matrix is augmented with a multiple of the identity, \( \alpha I \), to stabilize its inversion (this is required because of the very small sample size (12) relative to the dimension (6), and of some collinearities in the predicted properties where amino acids are the same in each sequence for any given position; \( \alpha \) is a tuning parameter that we set to the smallest value possible \( 1 \times e^{-2} \)). Computing \( d_i \) for every position along the protein produces a distribution with mean \( \mu_d \) and standard deviation \( \sigma_d \). The threshold for identification of structurally informative sites along the protein is set to

\[ \text{Threshold} = \mu_d + 2\sigma_d \]

Sites with distances greater than the threshold were identified as structurally informative sites (SInS). The distances were normalized across the protein sequence for each protein server and every value of within and between group sequence variability using a z-score calculation.

\[ Z_i = (d_i - \mu_d) / \sigma_d \]
where $d_i$ is the value of distance at an amino acid position, $\mu_d$ is the mean distance of all amino acid positions in the protein and $\sigma_d$ is the standard deviation for distance of all amino acid positions in the protein.

In this study SInS were identified between the groups Ga and Gb for each of the five servers, and then for each value of between group variability. We computed two different indices to test how different servers and sequence variability levels affected our ability to identify SInS. Sensitivity was computed as the ratio of the number of sites identified as SInS over the total number of sites for each analysis (56, which was the sequence length). Specificity was computed as the ratio of SInS identified over the total number of true SInS (SInS that were known to be different between Ga and Gb from literature). We expected 7 SInS, which were the amino acid positions identified by Alexander et al. as necessary to maintain differences in folds between Ga and Gb.

### 3.2.5 Coefficient of variation (CV)

The coefficient of variation was used to consider multiple distances from different prediction servers and distances from different values of predicted within and between group variability for the identification of SInS. The coefficient was calculated as follows:

$$CV_i = s_i/m_i$$

where $s_i$ and $m_i$ are standard deviation and mean of the distances for amino acid position I across the set under consideration. This coefficient was computed separately for two sets; prediction servers, and sequence variability levels.
3.3 Results

3.3.1 Prediction servers and SInS

We test five popular structure prediction servers to understand how differences in their underlying assumptions and algorithms affect identification of SInS between 4 Ga and 4 Gb sequences (Fig 3-1). As the prediction models used differ, the servers predict with different accuracy under different circumstances. We test how this variability in output from each server could affect identification of SInS using the predicted values in input to our pipeline. Fig 3-2 shows the distances between Ga and Gb across the sequence for each prediction server. The true SInS are highlighted with dots. From Fig 3-2 we see the true SInS all have high distances. In terms of sensitivity, we identified anywhere from 2 – 6 sites as SInS based on predicted structural properties from the different servers (Table 3-2). In terms of specificity, out of the sites identified as SInS we found that ExPASy, Psipred and Predator did not identify any false positives (Table 3-2). Jpred and Predict Protein on the other hand showed false positive with Jpred having the largest number (Table 3-2) as the number of correctly identified SInS (Specificity) was lower than the number of identified SInS (Sensitivity). The other three servers showed equal specificity and sensitivity which meant that all identified SInS for these servers were correctly identified. None of the prediction servers were able to provide data enough to identify correctly all 7 true SInS. This was in part due to erroneous structure prediction and in part also due to the stringent cutoff we used to demarcate SInS (In some sets the correct positions show peaks that are not large enough to be over the imposed cutoff).
Figure 3-2: The normalized distances (y-axis) between Ga and Gb plotted across the amino acid position (x-axis) of the protein for each of the five secondary structure prediction servers used. The positions of true SInS are marked with a circle

Table 3-2: Sensitivity (no. of identified SInS over the length of the protein) and Specificity (no. of correctly identified SInS over no. of true SInS) values for each prediction server

<table>
<thead>
<tr>
<th>Server</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expasy</td>
<td>4/56</td>
<td>4/7</td>
</tr>
<tr>
<td>Psipred</td>
<td>4/56</td>
<td>4/7</td>
</tr>
<tr>
<td>Jpred</td>
<td>6/56</td>
<td>2/7</td>
</tr>
<tr>
<td>Predict Protein</td>
<td>2/56</td>
<td>1/7</td>
</tr>
<tr>
<td>Predator</td>
<td>4/56</td>
<td>4/7</td>
</tr>
</tbody>
</table>
3.3.2 Sequence variability and SInS

We tested the effect of within and between group sequence variability on the identification of SInS. Prediction servers like ExPASy predict probability scores based on a combination of amino acids in a sliding window [74]. Multiple combinations can give rise to a prediction for an alpha helix although the value of the prediction can vary in terms of absolute numbers. Such small change in values within a group can affect group averages in our calculations. We tested a wide range of within group and between group sequence variability (Table 3-1). Fig 3-3 shows the distances between the 12 different groups of Ga and Gb sequences (each group has 10 Ga and 10 Gb sequences and a different values for between within and between group variability) across the amino acid positions. The 12 pairs were obtained from combinations of 4 groups of Ga (A1 – A4) and 3 groups of Gb (B1 – B3) sequences. We saw that irrespective of sequence variability there was little difference in values of distance across the sequence. The sensitivity for the different pairs was quite similar with 3 – 4 sites being identified as SInS in all cases (Table 3-3). The specificity showed that there were no false positives detected in any case (Table 3-3). Also here no single comparison could identify all 7 true SInS mostly due to the stringent cutoff we imposed to remove false positives.
Figure 3-3: The normalized distances (y-axis) between Ga and Gb plotted across the amino acid position (x-axis) of the protein for different values of between group variability. Four groups of Ga (A1 – A4) sequences and three groups of Gb (B1 – B3) sequences are used to identify SInS between twelve different pairs of Ga and Gb. The positions of true SInS are marked with a circle.
**Table 3-3:** Sensitivity (no. of identified SInS over the length of the protein) and Specificity (no. of correctly identified SInS over no. of true SInS) values for different within and between group variability.

<table>
<thead>
<tr>
<th>Combination</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1B1</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A1B2</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A1B3</td>
<td>4/56</td>
<td>4/7</td>
</tr>
<tr>
<td>A2B1</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A2B2</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A2B3</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A3B1</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A3B2</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A3B3</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A4B1</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A4B2</td>
<td>3/56</td>
<td>3/7</td>
</tr>
<tr>
<td>A4B3</td>
<td>4/56</td>
<td>4/7</td>
</tr>
</tbody>
</table>

**3.3.3 Coefficient of variation**

In both our analyses (effect of prediction servers and of sequence variability) we were not able to identify all 7 true SInS. We used the coefficient of variation which computes variation at each amino acid position “across an analysis” to increase our sensitivity to identify all possible SInS. The basic idea behind using the coefficient of variation is that at true SInS we would expect low coefficient of variation (as distance values would be similar irrespective of differences in structure prediction) and high
distance (at the respective amino acid position). We took the normalized distances computed for all five prediction servers and at each amino acid position calculated the corresponding coefficient of variation. We plotted the coefficient (y axis) against distances (x axis) computed from the data predicted by the Jpred server (Fig 3-4a). For the sake of illustration we used data from the Jpred server as it had the maximum number of false positives identified as SInS. We wanted to test with this data whether the false positives could be separated from the true positive SInS. The data points on this XY plot was further clustered to identify the true SInS. Hierarchical clustering was implemented in R using the hclust function from the stats package [54]. From Fig 3-4a we see that the seven true SInS cluster neatly (marked by red circle in the bottom right quadrant), have very low coefficient of variation and can be separated from the false positives (green circle in the top right quadrant) that have high distance but also higher coefficient of variation. The same coefficient was computed across different sequence variability levels. Fig 3-4b shows that when the coefficient was plotted against distances from any pair of Ga and Gb (all 12 pairs were tested, A1B1 shown as an illustration), we could identify clearly the 7 true SInS (red circle). Thus the coefficient of variation helped identify SInS across multiple sets of data. This made our method more robust to identification of SInS between two groups of protein sequences when we had multiple sets of distance values for the same query sequence.
Figure 3-4: The CV (Coefficient of variation) on the vertical axis plotted against normalized distance between Ga and Gb on the horizontal axis. A) The distances used were between Ga and Gb calculated using data from Jpred server B) Distances used were between A1B1 for all amino acid positions. The bottom right corner has sites with high distances and low CV. The red ellipse highlights the true SInS in both cases. The green ellipse in Fig 3-4A shows the false positives with high distances and high CV.
3.4 Discussion

Prediction servers all use different methods to process a query sequence to predict secondary structure properties. Here we develop a statistical pipeline to predict the positions of structural differences between related groups of proteins and test the effects of different methods of secondary structure prediction and different levels of sequence variability on this pipeline. As test set, we used sequences that already have structures solved in PDB [8]. We find that servers based on Chou Fasman/GOR (ExPASy), Nearest Neighbor (PREDATOR) and a combination of Neural Network and Nearest Neighbor (PSIPRED) works the best for our dataset (Table 3-2) [18,66,68,69,70,71,72,73,74]. These servers allow us to identify the same 4 out of 7 true SInS in the dataset, without identifying a single false positive. The servers Jpred and Predict Protein perform quite poorly in this analysis. Jpred identifies 4 sites incorrectly as false positives, and only 2 sites out of 7 correctly, whereas Predict Protein identifies one site correctly and one site as false positive. Although the structure exists for these sequences we think the training sets used in these two servers are not broad enough to predict correctly the structure associated with each sequence. It is important to note that our results do not mean that servers employing neural networks are inefficient in general. They only suggest they may be for the proteins considered here. The main point to take away from this study is that different servers with their underlying methods and assumptions do in fact affect our ability to identify SInS. The three servers that allowed us to identify four SInS correctly also showed high distances between the two groups at the other three true SInS (Fig 3-2). These sites are not picked up in our analysis because they failed to have distances above
the threshold. This suggests that our threshold in these cases is so stringent that we get false negatives in our result but on the other hand this stringency helps us eliminate false positives.

We do not find marked effects of sequence variability on identification of SInS. In Fig 3-3 all 7 SInS have high distances between the groups (Ga and Gb), and all 12 pairs of Ga and Gb tested here show nearly identical distances across the sequence. This suggests that all the sequence substitutions in these cases are correctly identified as having no effects on the overall structure. In all the 12 pairs of Ga and Gb tested we are able to identify 3 - 4 sites as SInS out of the possible seven true SInS. This is again due to the stringent threshold implemented in the approach.

When several related datasets are available, we can use the coefficient of variation to identify true positives from the overall data. The idea behind this is that true positives will always be identified with higher distance, and hence will have lower variation as compared to false positives that do not have high distance in all the datasets. In the study presented here the CV helps us identify correctly all the 7 true SInS in both the cases. This study shows strengths and weaknesses of the statistical pipeline developed. We observe how the pipeline can be affected by variability due to different servers, but is robust within a broad sequence variability range. We also observe that when data across multiple servers based on the same input sequences is available, one could improve identification of true SInS using the coefficient of variation.
In the next chapter we study the structural evolution of proteins between human and avian metapneumovirus. We test how our statistical pipeline fares in identifying SInS from sequences of protein whose structures are not present in the protein database.
Chapter 4

Predicting secondary structural properties that distinguish the proteins of human and avian Metapneumovirus

4.1 Introduction

Populations that are separated by geographical or ecological barriers accumulate substitutions in gene sequences that allow for phylogenetic group separation [76,77,78]. Some substitutions that change amino acid sequence become fixed in populations because they affect fitness [4,26]. However, there is no direct correlation between change in sequence and structure [7]. Thus, while sequence differences suggest adaptation, they do not provide information about structural relatedness of phylogenetically defined groups [5,6].

The tertiary structure of most proteins remains unknown and challenging to predict [79,80], notwithstanding recent increases in the pace at which 3-dimensional structures are being solved. In contrast, secondary structural properties of alpha helix, beta sheet and coil, and biophysical properties including hydrophobicity, accessibility, and flexibility, reflect local interactions among residues and can be estimated with some confidence from the primary amino acid sequence [67,74,81]. Structural elements can be conserved even in the absence of sequence similarity, as multiple combinations of amino acid sequence can give rise to similar secondary structural properties [6], and certain structural features, for example alpha helices, are more often found in regions of low sequence conservation than are beta sheets [24]. Moreover, secondary structural features
have been shown to be involved in adaptive changes in protein structure. For example, Colins et. al. [26] demonstrated that glycosidases enzyme of organisms that have adapted to higher temperatures have a decrease in values of predicted flexibility compared to organisms adapted to lower temperatures.

Changes at sites in the viral protein sequence that provide adaptive advantage in a new host come under positive selection [7]. Measures of positive selection are therefore often used to identify adaptive changes in evolving virus populations. However, using positive selection to infer structural changes in proteins has drawbacks. First, positive selection at a site is usually calculated as a ratio of non-synonymous (dn) to synonymous (ds) substitutions. This ratio of dn/ds has been shown to decrease over longer divergence times due to saturation at non-synonymous sites and accumulation of synonymous substitutions at any given position [12]. Second, non-synonymous substitutions can also occur by stochastic processes at sites that are not functionally constrained and offer no adaptive advantage [82]. Finally, as noted above, secondary structural features are determined by local properties and can be due to substitution over a region of the protein instead of at a single site.

Viruses can establish successful infections in new host species and adapt rapidly to their new environment. It is unclear if adaptation to a new host is accompanied by changes in secondary structure of proteins. The statistical pipeline developed in Chapter 3 was used to identify the sequence locations of secondary structure differences between two related groups of viruses that have adapted to different hosts. We use as a model metapneumovirus adapted to avian and human hosts. Metapneumovirus is a respiratory
pathogen of the Paramyxoviridae that was first identified as a human pathogen in 2001 [83], and is believed to have originated in birds, crossing the species barrier within the last 200 years [84]. This is a negative-sense, non-segmented RNA virus consisting of nine coding regions (N-P-M-F-M21-M22-SH-G-L) with total genome size of 15Kb [31], and has been reported to be among the faster evolving RNA viruses [84,85,86]. There are full length genome sequences of metapneumovirus available to identify changes in predicted structure between the avian and human metapneumovirus groups.

The goal of this study was to use the previously developed statistical framework to identify amino acids that confer different predicted secondary structural features to the proteins of two phylogenetically related groups of viruses, the avian and human metapneumovirus. We used a multivariate dataset comprising six predicted secondary structural or biophysical properties (alpha helix, beta sheet, coil, hydrophobicity, flexibility and accessibility) and quantified a distance between avian and human metapneumovirus proteins at each amino acid position based on these properties. We identified positions that differentiated the avian and human metapneumovirus protein groups based on the six predicted structural properties. We also showed that the number of sites conferring differences in predicted structure do not correlate with the number of changes in the amino acid sequence between avian and human metapneumovirus. The sites that were identified to differ in predicted secondary structural properties retained signatures of phylogenetic group affiliation, with only 1 out of 49 exhibiting evidence of positive selection. Further analysis demonstrated that these sites accumulated synonymous changes at the same rate as the rest of the protein, suggesting that these sites
appear early during adaptation to the host. We also determined that most of the properties conferring distance between the two groups of viruses affected thermodynamic stability of the proteins.

4.2 Material and Methods

4.2.1 Data and phylogenetic sequence analyses

Twelve full length metapneumovirus genomes including six human (Accession: AF371337, AY297748, AY207749, AY525843, DQ843658, DQ843659), five avian subtype C (Accession: AY579780, AY590688, DQ009484, EF199771, EF199772), and one avian subtype A isolates (Accession: AY640317) were retrieved from GenBank [87] for this study (www.ncbi.nlm.nih.gov/genbank/). Individual protein sequences were aligned using ClustalW [88]. Prior to the structural analyses, we checked the phylogenetic congruency among the nine genes by performing the Shimodaira-Hasegawa test [89] implemented in PAUP* 4.0 [90]. Maximum likelihood trees, with the appropriate nucleotide substitution models selected by the Akaike Information Criterion method implemented in Modeltest ver. 3.7 [91], were reconstructed using PAUP* 4.0. Using the RDP3 program [92] sequences were tested for recombination. None of the sequences had evidence of recombination based on any criteria. The number of polymorphic and phylogenetically informative sites (PInS) was calculated for each protein using MEGA version 4 [93]. Polymorphic sites were identified as sites with at least two different amino acids in the alignment. Phylogenetically informative sites were identified as polymorphic sites for which the alignment contained at least two amino acids, each present at least twice. To identify codons under positive selection, we
performed a codon-by-codon selection analyses for each protein coding region using the Fixed Effects Likelihood (FEL) method [94,95] available on the Datamonkey webserver (www.datamonkey.org). The selection analysis for each protein was performed using the appropriate nucleotide substitution model selected for that protein by the Modeltest program available on the Datamonkey webserver.

4.2.2 Identification of structurally informative sites

The twelve full genome sequences used in the phylogenetic analyses were also used to identify differences in protein predicted secondary structures between the two groups of metapneumovirus. Six structural properties, hydrophobicity (Kyte and Doolittle [96]), accessibility (Janin [97]), flexibility (Bhaskaran and Ponnuswamy[98]), alpha helix (Deleage and Roux[75]), beta sheet (Deleage and Roux[75]) and coils (Deleage and Roux[75]) were predicted for each amino acid site along the sequences of each protein using the ExPASy Proteomics Server [74,81], (www.expasy.ch/tools/protscale.html). In terms of tuning parameters for this algorithm, we used a window size of 5, and an exponential weight for each window (with the window edges at 10% weight). The six values in output for each position (predicted properties) were normalized to range between 0 and 1.

After aligning the proteins’ amino acid sequences with ClustalW [88], twelve 6-dimensional vectors (based on the six structural properties predicted by ExPASy; one such vector for each sequence) were associated with every un-gapped alignment column (i.e. site) along each protein. Positions with gaps were excluded from further analysis because structural properties cannot be associated with gaps. The exclusion of amino
acids with gaps creates a discontinuous horizontal axis when values are plotted across the protein. Next, we used the statistical pipeline developed in chapter 3 to compute distances between the avian and the human metapneumovirus at each amino acid position. To compute the coefficient of variation (CV), we randomly sub-sampled 4 sequences each from the avian and human metapneumovirus groups to compute the value of distance at each position. This was repeated 225 times (for all possible combinations of 4 avian and 4 human metapneumovirus sequences) and CV was computed across the 225 values of distances at each amino acid position based on the pipeline outlined in chapter 3. Structurally informative sites (SInS) were identified based on hierarchical clustering on the XY plot containing CV and distance (between the six avian and six human metapneumovirus sequences).

Phylogenetic clustering for the concatenated structurally informative amino acid sites and for the concatenated amino acid sequences of the nine proteins was implemented using the distance based neighbor-joining method implemented in MEGA version 4 [93].

4.2.3 Synonymous substitutions at SInS

For each protein, we tested whether the rate of synonymous substitutions (ds) at SInS differed from that at other sites. We obtained ds value for each amino acid position along the protein from a FEL analysis implemented on the Datamonkey webserver [79, 80]. After computing mean and median ds for the N SInS in the protein, we drew 100,000 random samples of size N from the ds distribution across the protein, and computed mean and median for each such sample. These 100,000 means (medians) were treated as a null
distribution against which to compare the mean (median) computed on the N SInS. This allowed us to derive empirical p-values to test consistency of the mean (median) ds from N SInS with the overall mean (median) ds for the protein (e.g., the empirical p-value to test for consistency against the SInS mean being larger than the overall mean is obtained calculating what fraction of the 100,000 means is larger or equal to the SInS mean). We considered all p-values smaller or equal to 0.05 as significant.

4.2.4 Structural property pairs driving SInS

The direction along which avian and human sequences separate at each site is expressed as a linear combination of the six predicted structural properties used in the analysis. The coefficients in this combination allowed us to identify the two properties that drive the separation at the given site. In order to investigate any tendency of properties to interact in determining SInS, for the entire genome we analyzed the representation of each of the 15 possible pairs of drivers at these sites. For the whole genome, the expected frequencies of pairs under an “absence of an interaction” null hypothesis were calculated as the number of SInS in the genome over 15, and compared with the observed frequency of the pairs using a chi-square test. Significance was benchmarked at the 0.05 level, and a Yates correction was used because of the small sample sizes available for each comparison.

4.3 Results

4.3.1 Metapneumovirus sequence polymorphism

To determine relationships between sequence diversity and substitutions that could affect protein secondary structure in evolving viral proteins we first explored the
sequence diversity of the avian and human metapneumovirus genomes used in this study. Sequence diversity between these proteins was determined by estimating the percentage of polymorphic sites in the alignment of each protein. Polymorphic sites were defined as sites that had at least two different amino acids present at any position in an alignment, thus all polymorphic sites had a dn > 0. The percentage polymorphism in metapneumovirus proteins varied from 28.3% in the M protein to 90.4% in the SH protein; M22, SH, and G (which are contiguous proteins in the metapneumovirus genome) had the highest percentage polymorphism in their sequences (Table 4-1). Only a subset of polymorphic sites was phylogenetically informative (PInS) because an amino acid change in a single sequence in the alignment contributed to polymorphism but not necessarily to phylogenetic affiliation. Nevertheless, the percentage of PInS was directly correlated with the percentage of polymorphic sites and ranged from 13.7% in N protein to 72.9% in SH protein (Table 4-1).

Table 4-1. Evolutionary genetics and structurally informative sites of the proteins in metapneumovirus.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Polymorphica sites - proteins (%)</th>
<th>PInSb – protein (%)</th>
<th>SInSc (%)</th>
<th>Sites under positive selection (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>35.5</td>
<td>13.7</td>
<td>1.8(7)</td>
<td>0.0</td>
</tr>
<tr>
<td>P</td>
<td>50.7</td>
<td>36.1</td>
<td>2.7(8)</td>
<td>0.3</td>
</tr>
<tr>
<td>M</td>
<td>28.3</td>
<td>14.2</td>
<td>2.8(7)</td>
<td>0.0</td>
</tr>
<tr>
<td>F</td>
<td>37.0</td>
<td>21.5</td>
<td>1.1(6)</td>
<td>0.0</td>
</tr>
<tr>
<td>M21</td>
<td>32.0</td>
<td>16.5</td>
<td>1.1(2)</td>
<td>0.5</td>
</tr>
<tr>
<td>M22</td>
<td>78.1</td>
<td>46.6</td>
<td>4.2(3)</td>
<td>0.0</td>
</tr>
<tr>
<td>SH</td>
<td>90.4</td>
<td>72.9</td>
<td>2.2(4)</td>
<td>3.1d</td>
</tr>
<tr>
<td>G</td>
<td>75.7</td>
<td>34.6</td>
<td>2.0(6)</td>
<td>3.0</td>
</tr>
<tr>
<td>L</td>
<td>41.8</td>
<td>22.2</td>
<td>0.3(6)</td>
<td>0.2</td>
</tr>
</tbody>
</table>

(a) Polymorphic sites contain at least two types of amino acid in an alignment.
(b) PInS are Phylogenetically Informative Sites, which are a special subset of polymorphic sites that provides phylogenetic grouping of the avian and human metapneumovirus groups.
(c) SInS are Structurally Informative Sites, which are a special subset of polymorphic sites that also confer predicted secondary structural changes between avian and human metapneumovirus groups. Both percentages and absolute numbers (in brackets) of SInS are shown here.
(d) A single SInS is also identified as a positively selected site in the protein.
4.3.2 Identification of structurally informative sites

Using the pipeline described in chapter 3 we identify amino acid positions in the alignment of each protein associated with secondary structure differences between avian and human metapneumoviruses. Our analyses revealed that 0.3 to 4.2% of amino acid positions in each protein of the metapneumovirus were structurally informative (SInS; Table 4-1). At 43% of these sites, the amino acid was identical within all avian sequences and within all human sequences but differed between the two groups. The remaining 57% of SInS were polymorphic at the amino acid level within either the avian or the human metapneumovirus groups. Of these within-group polymorphic sites, 14% were polymorphic in the human metapneumovirus sequences. Polymorphism at SInS within either the avian and/or the human protein sequence suggested that nonsynonymous substitutions at that site did not affect the properties that led to the elevated distance between the avian and human metapneumovirus groups. As we average over the two groups while calculating distance, any large change in predicted secondary structural properties arising due to within group nonsynonymous substitution could increase within group distance and impact the distances between the groups. The fact that polymorphic SInS were identified suggests that amino acid variation within the group conserved secondary structural properties.

The amino acid positions of SInS for G and F proteins and the corresponding XY plots of CV and distance are shown in Fig 4-1. For the F protein, 6 sites (1.1% of the total sites) were classified as SInS. Distances at sites other than these peaks were low, indicating that secondary structural properties were conserved in these regions of the
metapneumovirus F protein between avian and human groups (Fig 4-1a). In contrast distances in the G protein were elevated across most of the protein but there were still only 6 SInS (2.0% of the total sites) identified (Fig 4-1b). The high distances are due, in part, to high sequence polymorphism in the G protein (Table 4-1).

**Figure 4-1:** Structurally informative sites (SInS) across the length of the membrane associated glycoproteins F (A) and G (B). The horizontal axis in both graphs indexes amino acid positions along the glycoproteins. Residue number on the horizontal axis is discontinuous as some amino acids are omitted in the analysis due to gaps in the alignment. The blue line represents the distance in predicted secondary structure properties between avian and human metapneumovirus groups at each amino acid position, and refers to the right vertical axis. Red dashed bars identify the position of SInS that have been identified by a combination of coefficient of variation (CV) and hierarchical clustering, and refer to the left vertical axis – which ranges from 0 (non informative sites) to 1 (SInS). The identified SInS are circled in red on the respective CV versus distance XY plots (C and D). The horizontal axis (X) for both graphs refers to the values of distance and the vertical axis (Y) indexes the corresponding CV for both proteins. The SInS identified after hierarchically clustering the dataset are circled in red in each protein.
4.3.3 Comparison of polymorphic sites and structurally informative sites

There was no observed association between percentages of polymorphic sites or PInS and percentage of SInS for any of the metapneumovirus proteins (Table 4-1). PInS correlated strongly with polymorphic sites (Fig 4-2a), but SInS did not (Fig 4-2b). The proteins that were most polymorphic in nature (M22, SH and G) had similar percentages of SInS as compared to the rest of the proteins of the metapneumovirus (Fig 4-2b). This suggested that most amino acid replacements in the sequence did not induce changes in secondary structural elements between avian and human metapneumovirus groups.
Figure 4-2: Relationship between percentages of phylogenetically informative sites (PInS) and polymorphic sites (A), and between percentages of structurally informative sites (SInS) and polymorphic sites (B), across the proteins of the metapneumovirus. The percentages are mapped on the vertical and horizontal axes of the scatter plots, and the proteins represented as color-coded points – along with least square lines and the corresponding shares of variability explained (R²). PInS are sites in an alignment where at least two amino acids occur, each with a frequency of at least two. SInS are sites in an alignment that differ significantly in secondary structural properties between the avian and human metapneumovirus groups.

To determine the extent of phylogenetic information that was contained in SInS, we compared neighbor joining trees based on concatenated protein sequences with those estimated from concatenated SInS (Fig 4-3). The tree based on SInS alone maintained the same configuration of human A and human B metapneumovirus as that estimated from
the full amino acid sequence data (Fig 4-3b). This suggested that polymorphism at SInS within the human metapneumovirus group was responsible for the phylogenetic clustering of the human A and human B metapneumovirus subgroups (Fig 4-3b). Since these sites were identified as SInS we expected within group sequence changes to be neutral to changes in structure within the group.

Figure 4-3: Neighbor joining trees showing the phylogenetic relationships among different subtypes of metapneumovirus isolated from human and avian hosts. The first tree (A) is based on the concatenated protein sequences, and the second (B) on the concatenated SInS for the whole genome.
4.3.4 Characteristics of structurally informative sites

Twenty-nine sites were identified to be under positive selection in the genome. Of these, only one was identified as SInS in the SH protein (Table 4-1). Interestingly, the G protein had the largest number of sites under positive selection (18) but none that were identified as SInS (Table 4-1).

It is possible that SInS arose early during adaptation of metapneumovirus to the avian and human hosts and are not identified as being under positive selection because they have accumulated synonymous changes at the same rate as other sites in the metapneumovirus genome. Since positive selection is calculated as a ratio of non-synonymous to synonymous substitutions, accumulation of synonymous substitutions could obscure identification of a site as positively selected. On the other hand, if the SInS arose recently, we would expect these sites to have lower ds compared to the rest of the protein. This was tested by comparing observed ds averages at SInS to appropriate null distributions obtained by re-sampling sites. The resulting empirical p-values were non-significant (> 0.05) for all of the metapneumovirus proteins, which suggested that in fact SInS arose early during the adaptation process.

At each structurally informative site, we identified the two predicted properties (out of six) that maximally contributed to the distance between the avian and human metapneumovirus groups. The frequencies with which these protein properties appear together at SInS across the genome appeared to be non-random, indicating systematic associations. We assessed these associations using chi square tests. The pair of alpha helix and coil occurred more frequently than expected by chance (p < 0.05) across the
entire metapneumovirus genome, and these two properties were always negatively correlated. This suggested that at these SInS, one group had a predicted structured alpha helix and the other a less ordered coil. This tradeoff in properties between the two groups could possibly affect the overall packing and thermodynamic stability of the proteins of the metapneumovirus.

4.4 Discussion

The host specific adaptation of viruses is accompanied by amino acid substitutions in their genome. However, it is unclear how amino acid substitutions that occur during adaptation to a new host are reflected in changes in protein structure, as changes in protein sequence do not always correlate with change in structure [5,6,99]. Here we determined if there were sites that could distinguish avian and human metapneumovirus proteins based on predicted secondary structural properties. This approach complements analyses based on primary amino acid sequence because secondary structure properties reflect regional attributes conferred by individual amino acids [6]. Extensive sequence polymorphism does increase distances at all sites (Fig 4-1b) because groups of amino acid contribute differently to the same structural properties. In six dimensions such differences can result in distances between the two groups. The statistical pipeline developed in chapter 3 allowed us to identify positions with significant predicted structural differences between the avian and human metapneumovirus groups even when there was polymorphism at the sites within the groups.

We demonstrated that there is little association between amino acid diversity and differences in predicted secondary structure properties in the avian and human
metapneumovirus proteins. Illergard et al., [6] have shown that protein structures among even distantly related eukaryote groups are more likely to be conserved than their corresponding nucleotide and amino acid sequence. We found that the relationship between amino acid diversity and differences in predicted secondary structural properties varied among metapneumovirus proteins. For example, there were 20 times more polymorphic sites than SInS in the F protein, but the G protein revealed more than 38 times higher percentages of polymorphic sites compared to SInS (Table 4-1). Thus, few polymorphic positions in G conferred differences in predicted secondary structural properties between these avian and human metapneumovirus proteins. In contrast, the F protein was more conserved at the amino acid level but harbored a higher proportion of amino acid substitutions contributing to predicted secondary structural differences between the two metapneumovirus groups. Amino acid polymorphism within the avian or human metapneumovirus groups at SInS suggested that differences in predicted secondary structure could be identified between the two groups of metapneumovirus in the presence of amino acid diversity within groups. This is consistent with the lack of direct association between primary sequence and structure, which has been described [5,6,99], and highlights the challenges involved in studying structural changes in proteins based on amino acid sequence.

Viruses are believed to adapt to a new host in part under strong selection pressure from the host immune system [28]. Viral epitopes change rapidly under immune pressure, often through substitutions that prevent interaction with receptors on host B and T cells and can arise with no or small associated structural change [27,29,30,100,101]. The genes
under strongest positive selection in the metapneumovirus are the G and SH, which show fewer SInS than the other proteins. We reason that this may be in part due to immune evasion, as the G and SH proteins are known to contain immunogenic epitopes [102]. Our results also suggest that SInS are fixed early in the adaptation process because in our study, there is evidence that these sites accumulate synonymous changes at the same rate as the rest of the genome. Hence many SInS are not identified as positively selected sites.

The secondary structural features that we evaluated contribute to protein stability. A change from a coil to more structured sheets and helices can also increase stability in proteins [4,103]. The most common type of structural change observed in our data is defined by the inverse correlation of the alpha helix and coil “driving pair”, indicating that a change had occurred from less ordered coils to more stable helices in the avian and the human metapneumovirus proteins. These associations may be driven in part by temperature differences in the two host systems. The avian mean body temperature is 40°C whereas that of humans is near 37.8°C [104]. Changes in temperature can lead to altered stability in homologous proteins [105]. Our findings indicate that host body temperature may contribute to the adaptive evolution of viruses in human and avian hosts. These are important outcomes of our analysis that would not be discernable using phylogenetic methods and are worthy of further investigation because important viral infections of humans emerge from avian hosts. A better understanding of potential structural changes that consistently arise in viral proteins, most of which do not have a crystal structure, during adaptation to a new host could inform novel therapeutic strategies for emerging viral infections.
In the next chapter we study the evolution of structure reverse transcriptase of primate lentiviruses. We test how identified SInS are informative to changes in tertiary and quaternary structures of a protein.
Chapter 5

Identification of predicted structural changes in Reverse Transcriptase within the primate lentivirus lineage

5.1 Introduction

Lentiviruses are retroviruses that are associated with a wide range of diseases in mammals. With the discovery of human immunodeficiency virus (HIV) in the 1980s lentiviruses have become a focus of study worldwide. Lentiviruses are primarily divided into five phylogenetic groups that correspond to the host species, which include immunodeficiency viruses of ovine (sheep, goat), bovine (cattle), equine (horses), feline (cats), primates (monkeys, humans) [106]. The genome of these viruses are single stranded positive sense RNA (present in two identical copies) which are transcribed to double stranded DNA inside the host cell by the viral reverse transcriptase (RT) protein [107]. The RNA of the HIV-1 genome consists of nine genes [108]. There are two known human immunodeficiency viruses, HIV1 and HIV2. These viruses cause a decline in the CD4+ T cells in the body leaving the host susceptible to other opportunistic pathogens [108].

Within the primate lentivirus group the major hosts for the viruses are humans (HIVs) and non human primates (SIVs). SIV is known to infect at least 33 [109] different species of non human primates and some well-studied SIVs are SIVsm (sooty mangabeys), SIVmac (macaques), SIVagm (African green monkeys), SIVmnd (mandrills), and SIVcpz (chimpanzees) [110]. A phylogeny based on the sequences of primate lentivirus shows that the human viruses (HIV1 and HIV2) were introduced into
the human population at two independent times from SIVcpz and SIVsm, respectively [110]. These two events therefore represent cases of independent evolution of a primate lentivirus in the human host. The primate lentiviruses evolve at a very rapid rate, with HIV1 having a mutation rate of around $3.4 \times 10^{-5}$ nucleotide changes per site per replication cycle [28,111]. This is primarily due to the viral reverse transcriptase protein that lacks the proofreading activity, thus incorporating errors in the viral genome during replication [32]. The fidelity of reverse transcriptase differs between the members of the primate lentiviruses, with HIV1 showing the highest error rates [32,33,108,111,112].

The lentiviral RT is a functional heterodimer consisting of two subunits the p66 and the p51 [32]. The p66 and the p51 are coded from the same region of the pol gene with the p66 having an extra RNAse H domain at the C terminus that is cleaved in the p51 subunit. The p66 is 560 amino acids in length whereas the p51 is ~430 amino acids long [32,34]. The two subunits share the same sequence but they fold very differently. The polymerase and the RNAseH active sites of RT are present on the P66 subunit. The polymerase domain can be further broken down to four different sub domains, fingers (amino acid 1 – 85, 118 – 155), palm (amino acids 86 – 117, 156 – 236), thumb (amino acids 237 – 318) and connection (amino acids 319 – 426) [32,34]. The RNAseH domain of RT that is absent from the p51 subunit extends from amino acid 427 up to the end of the protein. RT is shown to be functional mostly as a heterodimer [113,114]. The stability of the dimer interface also differs between the primate lentiviruses, with the HIV1 having the weakest interaction at the dimer interface [114]. Drugs are targeted to destabilize the heterodimer to prevent its formation, thus rendering the protein and virus ineffective.
Thus, identifying structural differences (both secondary structure differences, SInS and how they interact on a crystal structure) between RT of the four lentivirus groups can help tease out the structural differences that may be responsible for changes in function e.g. differences in fidelity of the RT proteins.

In this chapter, we identify the predicted structural differences of RT in the primate subgroup using the statistical pipeline developed in chapter 3. The RT protein has a rich dataset of protein sequences in GenBank, with multiple crystal structures of HIV1 and one structure of HIV2 available in Protein Data Bank (PDB). In this project we studied the independent cases of RT evolution from SIVcpz to HIV1 and SIVsm to HIV2. The main question addressed in this chapter is whether predicted secondary structures in RT converged during adaptation of different primate viruses to humans. Our results indicated that although structurally informative sites (SInS) could be identified in the two evolutionary lineages, these did not fall at the same amino acid positions in HIV1 and HIV2 RT suggesting an absence of structural convergence at the amino acid level. It is possible still to maintain differences in structure by mutating different amino acids of the same structural element. The second question was to identify structural differences that arose in the RT of HIV1 and HIV2, which represent two viruses with independent evolution and adaptation in humans. We identified SInS that highlight structural changes that differentiate between the RT of HIV1 and HIV2. These sites may in part help understand the differences in fidelity and dimer stability of the two RTs [32,33,108,111,112].
5.2 Material and Methods

5.2.1 Sequence Dataset

Forty five sequences of RT from four different primate lentivirus groups were retrieved from GenBank for this study [87]. Out of the forty five sequences twelve sequences were of HIV1 (CAB96338, CAA06810, AAA44860, AAA99879, AAA44198, AAA45366, AAB50259, AAF18397, AAD03184, AAD17757, AAC29060, AAB36501) twelve of SIVcpz (AAF18575, AAM34554, AAN18271, ABD19484, ABD19502, ABD39700, ABD39703, ABQ51060, ABQ51069, ABD19475, AAO13960, ABD19493), twelve of HIV2 (AAB00737, AAB01352, AAB00764, AAA43942, AAC54467, P15833, 1MU2A, BAH97704, ABC39618, ACH73021, ABV83026, AAR98760) and nine of SIVsm (AAC56559, AAC68656, AAK55275, ABA54155, ABA54164, ADI24349, P12502, P12505, AAA47753). The sequences were chosen to sample all the various phylogenetic subgroups within the four groups studied (HIV1, HIV2, SIVcpz, and SIVsm). This was done to maximize the diversity of the RT sequences within any of the four given groups. The protein sequences were aligned using ClustalW [88].

5.2.2 Structural Properties

Secondary structural properties were predicted using the ExPASy webserver. The six predicted structural properties used in this study were hydrophobicity (Kyte and Doolittle [96]), accessibility (Janin [97]), flexibility (Bhaskaran and Ponnuswamy [98]), alpha helix (Deleage and Roux [75]), beta sheet (Deleage and Roux [75]) and coils (Deleage and Roux [75]). For the ExPASy webserver the tuning parameters used in the input form were as follows; the window size was set to 5 and an exponential weight for
each window (with the window edges at 10% weight) was used [74]. The six predicted values in the output for each amino acid position (predicted properties) were normalized to range between 0 and 1.

To predict the quaternary structural properties (predicted properties of the dimer interface) of RT the following pipeline was used. The RT protein sequences of SIVcpz and HIV1 were modeled using an existing HIV crystal (1rtj) and the SIVsm and HIV2 were modeled on the only available HIV2 crystal (1mu2) in the DeepView viewer [115]. The query sequence was first uploaded to DeepView followed by the template crystal structure. The sequence was then threaded and aligned to the template structure. This alignment file was submitted to the SWISS-MODEL to produce a prediction of the heterodimeric structure [115]. The output from SWISS-MODEL was saved as a pdb file that contained the structure and coordinates of the modeled protein. This pdb file was then uploaded to the PISA webserver to predict quaternary structural property of the dimer interface [116]. From the output of the PISA webserver the following predicted properties were used in this study; no. of atoms at the interface (p66), no. of residues at the interface (p66), no. of atoms (p51), no. of residues (p51), interface area, gain in solvation energy, p-value for the gain in solvation energy, no. of hydrogen bonds, no. of salt bridges, no. of disulphide bonds [116].

5.2.3 Statistical Analysis

After aligning the RT proteins’ amino acid sequences with ClustalW [88], twenty four 6-dimensional vectors (based on the six structural properties predicted by ExPASy; one such vector for each sequence) were associated with every un-gapped alignment
column (i.e. site) along each protein. Positions with gaps were excluded from further analysis because structural properties cannot be associated with gaps. The exclusion of the amino acids from the analysis at gaps created a discontinuous horizontal axis when values were plotted across the protein. Next, we used the statistical pipeline developed in chapter 3 to compute distances between RT of four pairs of primate lentiviruses (SIVcpz with HIV1, SIVsm with HIV2, HIV1 with HIV2, and SIVcpz with SIVsm).

Coefficient of variation (CV) was computed for each amino acid position for all four different comparisons (SIVcpz with HIV1, SIVsm with HIV2, HIV1 with HIV2, and SIVcpz with SIVsm) of RT. To compute CV for any comparison we randomly sampled 4 sequences from each group to compute the value of distance at each position. This was repeated 1000 times and CV was computed across the 1000 values of distances at each amino acid position based on the pipeline outlined in chapter 3. Structurally informative sites (SInS) for each comparison were identified based on hierarchical clustering on the XY plot containing CV and distance.

For the analysis of the predicted quaternary structural properties we used the 10 predicted properties from PISA webserver of all 45 sequences used in the study [116]. We implemented the LDA analysis in R using the LDA function from the MASS [54] package to separate out the four primate lentivirus groups based on the 10 predicted properties from PISA. The first two eigenvectors from the LDA analysis was then plotted in R [54].
5.3 Results

5.3.1 SInS between the primate lentiviruses.

SInS were identified in RT between four distinct groups within the primate lentiviruses (SIVcpz, HIV1, SIVsm, and HIV2), reflecting two independent cross over events of the virus into the human population. HIV1 shares a common ancestry with SIVcpz whereas HIV2 shares ancestry with SIVsm [32,108]. Identifying SInS in RT protein between HIV1/SIVcpz and HIV2/SIVsm would provide information about the independent structural evolution of primate lentivirus RT in humans. The numbers of SInS identified between the HIV1/SIVcpz is shown in Table 5-1. Between HIV1 and SIVcpz RT we identify 7 SInS, two of which lie at the dimer interface (Table 5-1). The SInS identified that differ between HIV1 and SIVcpz RT are highlighted in Fig 5-1A.

Table 5-1: Number of SInS between primate lentivirus groups

<table>
<thead>
<tr>
<th>Comparison</th>
<th>SInS Total</th>
<th>SInS at RT Dimer</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV1/SIVcpz</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>HIV2/SIVsm</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>HIV1/HIV2</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>SIVcpz/SIVsm</td>
<td>12</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 5-1: Structurally informative sites (SInS) across the length of the Reverse Transcriptase protein. (A) Between HIV1/SIVcpz and (B) Between HIV1/HIV2. The horizontal axis in both graphs indexes amino acid positions along the RT. Residue number on the horizontal axis is discontinuous as some amino acids are omitted in the analysis due to gaps in the alignment. The blue line represents the distance in predicted secondary structure properties at each amino acid position, and refers to the left vertical axis. Red dashed bars (in both A and B) identify the position of SInS that have been identified by a combination of coefficient of variation (CV) and hierarchical clustering, and refer to the right vertical axis – which ranges from 0 (non informative sites) to 1 (SInS). The green dashed bars in figure 5-2B highlights a subset of SInS between HIV1/HIV2 that is also present as differences between the ancestral SIV RTs.

We studied the distribution of the SInS identified in the RT proteins between HIV1 and SIVcpz, over the five distinct domains (finger, palm, thumb, connection, RNAseH) of RT (Table 5-2). There were no SInS in the finger or connecting domain of RT that distinguished HIV1 from SIVcpz suggesting that these domains had undergone no structural changes in HIV 1 when compared to the simian ancestors. When the SInS identified between HIV1 and SIVcpz RT were positioned on an HIV1 RT crystal we found SInS pairs that possibly interacted in three dimensions (criteria for interaction was distance less than 7Å between alpha carbon atoms). SInS at position 294 and 297 were
found to interact with both present at the top of the thumb domain (White circle Fig 5-2A). These positions (294 and 297) were also found to interact in the p51 subdomains. The change in properties at these positions suggests changes in accessibility and flexibility of the tip of the thumb. The HIV1 RT was more flexible than the ancestral SIV RT. Positions 452 and 467 (RNAseH domain) also possibly interact in the tertiary structure although they are discontinuous on the linear sequence. Both sites (452,467) show a change in the properties hydrophobicity and beta sheet suggesting better packing of the protein. The SIVcpz RT has higher values of hydrophobicity and beta sheet than the HIV1 RT suggesting a more stable packing in SIV.

**Table 5-2:** Percentage of SInS distributed in different domains of RT

<table>
<thead>
<tr>
<th></th>
<th>Finger</th>
<th>Palm</th>
<th>Thumb</th>
<th>Connection</th>
<th>RNAseH</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV1/SIVcpz</td>
<td>0.0</td>
<td>14.3</td>
<td>28.6</td>
<td>0.0</td>
<td>57.1</td>
</tr>
<tr>
<td>HIV2/SIVsm</td>
<td>0.0</td>
<td>0.0</td>
<td>25.0</td>
<td>50.0</td>
<td>25.0</td>
</tr>
<tr>
<td>HIV1/HIV2</td>
<td>47.3</td>
<td>10.5</td>
<td>31.6</td>
<td>5.3</td>
<td>5.3</td>
</tr>
<tr>
<td>SIVcpz/SIVsm</td>
<td>41.7</td>
<td>16.7</td>
<td>16.7</td>
<td>16.7</td>
<td>8.3</td>
</tr>
</tbody>
</table>

We also identified the predicted secondary structural properties that differed at each SInS and tested whether any secondary structural property were always overrepresented between HIV1 and SIVcpz RT. At each SInS, we identified the two properties most responsible for the distance between the two RTs. The distribution of the six predicted secondary structural properties across all the SInS of HIV1/SIVcpz RT shows that hydrophobicity and beta sheet are highly overrepresented in this case (Table
This suggests that the RT of HIV1 and SIVcpz differ from each other in packing and stability i.e. more energy is required to unfold the protein (lower delta G values).

**Table 5-3:** Percentage of secondary structural properties that differ at SInS

<table>
<thead>
<tr>
<th>Accessibility</th>
<th>Flexibility</th>
<th>Hydrophobicity</th>
<th>Alpha Helix</th>
<th>Beta Sheet</th>
<th>Coils</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV1/SIVcpz</td>
<td>7.1</td>
<td>7.1</td>
<td><strong>42.9</strong></td>
<td>0.0</td>
<td><strong>35.7</strong></td>
</tr>
<tr>
<td>HIV2/SIVsm</td>
<td>0.0</td>
<td>12.5</td>
<td><strong>50.0</strong></td>
<td>0.0</td>
<td><strong>25.0</strong></td>
</tr>
<tr>
<td>HIV1/HIV2</td>
<td>5.3</td>
<td>15.8</td>
<td>23.7</td>
<td>21.1</td>
<td>23.7</td>
</tr>
<tr>
<td>SIVcpz/SIVsm</td>
<td>4.2</td>
<td>12.5</td>
<td>33.3</td>
<td>16.6</td>
<td>25.0</td>
</tr>
</tbody>
</table>

We also compared the RT of HIV2 and SIVsm, another case of evolution and adaptation of the virus in the human host. We identified 4 SInS between the HIV2 and SIVsm RT protein with two sites present at the dimer interface. We studied the distribution of the SInS identified in the RT proteins between HIV2 and SIVsm, over the five distinct domains (finger, palm, thumb, connection, RNAseH) of RT (Table 5-2). There were no SInS in the finger and palm domain of RT that distinguished HIV2 from SIVsm suggesting that the finger and palm domain had undergone no structural changes when compared to the simian ancestors. None of the SInS identified between HIV2 and SIVsm RT interacted in the crystal structure (criteria for interaction was distance less than 7Å between alpha carbon atoms). We also identified the predicted secondary structural properties that differed at each SInS and tested whether any secondary structural property were always overrepresented between HIV2 and SIVsm. The structural changes across all the SInS between HIV2/SIVsm RT shows that
hydrophobicity and beta sheet are highly overrepresented in this case (Table 5-3). This suggests that the RT of HIV2 and SIVsm differ from each other in packing and stability i.e. more energy is required to unfold the protein (lower delta G values), which is also similar to differences observed between HIV1 and SIVcpz RT. It is important to note that no SInS were identified that were common (based on amino acid positions) between the two crossover events (HIV1/SIVcpz and HIV2/SIVsm RT) although some similarities were observed in the over representation of secondary structure drivers at the SInS (hydrophobicity and beta sheet).

We compared RT of HIV1/HIV2 to identify structural differences that arose between the two independent evolutionary events in humans. Between HIV1/HIV2 RT we identified 19 SInS with 4 sites lying at the dimer interface. We also compared the SIVsm/SIVcpz RT to identify how many SInS between HIV1/HIV2 RT out of the 19 identified were already present in the ancestral SIV RT sequences. There were 12 SInS identified between SIVxm and SIVcpz RT with 2 sites as the dimer interface (Table 5-1). When we compared amino acid position of SInS between the two HIVs (HIV1 and HIV2) and their ancestral SIVs (SIVcpz and SIVsm) RT we were able to identify 8 SInS that shared the same amino position in both the cases. These 8 SInS identify differences in predicted structure in the simian virus that have been maintained in the human hosts. Thus between HIV1/HIV2 RT there were only 11 unique SInS that have arisen in the two independent courses of adaptation to the human hosts (sites highlighted by red dashed bars in Fig 5-1B). These 11 unique SInS all lie in the finger, palm, thumb and connection suggesting that these four domains of HIV1 and HIV2 RT differed significantly based on
predicted secondary structure. No SInS are observed in the RNaseH domain between the HIV1 and HIV2 RT. In the thumb domain we could identify 3 sites (out of the 11 sites between HIV1/HIV2, sites 308, 313,315) that were different between the two HIV RTs but were absent in the ancestral SIV RTs (Table 5-2, Fig 5-2B). The sites clustered on the HIV1 crystal in both the p66 and the p51 subunits (distances between each amino acid are less than 7Å). This suggested differences in structure at the base of the thumb that has evolved in the two independent cases of crossover event of RT in humans (Fig 5-2B). The predicted structural properties at these three sites suggested a difference in flexibility between the two HIV RTs. We further tested whether any predicted structural properties were identified as overrepresented at these 11 SInS, but no such overrepresentation for any one property was observed (table 5-3).

Figure 5-2: SInS highlighted on the crystal structure of HIV1 RT (1RTJ) (A) Front (B) Back. The p66 subunit is colored blue whereas the p51 is shown in red. SInS highlighted in green were identified between HIV1/SIVcpz. SInS highlighted in white were identified between HIV1/HIV2. SInS highlighted in yellow were identified between HIV1/HIV2 but were present as differences between the ancestral SIV RTs. We can see clustering of SInS highlighted in white circles. The red circle highlights SInS that were identified only between HIV1/HIV2 RT at the base of the thumb domain.
5.3.2 Differences at the dimer interface

We used 10 different quaternary properties of the dimer interface from PISA webserver to differentiate between the four groups of primate lentivirus RT (HIV1, SIVcpz, HIV2, and SIVsm) [116]. When the first two eigenvectors of the LDA are plotted we find that the two lineages (HIV1/SIVcpz and HIV2/SIVsm) separate out completely but no clear clustering is observed within the two lineages (Fig 5-3). The two predicted quaternary properties that are responsible for the separation of the two lineages are the total numbers of residues/atoms interacting at the interface. In future work we would look at the SInS identified based on secondary structure and see whether they are responsible for changing the number of residues at the contact of the dimer interface. This would be information about quaternary structures gleaned from SInS identified based on secondary structures.
5.4 Discussion

In this chapter we study predicted secondary and quaternary structural differences between four primate lentivirus RT (HIV1, SIVcpz, HIV2, and SIVsm). The main question was to study how the HIVs have evolved from their ancestral SIVs in two independent cases of crossover events and whether there were signs of convergent evolution between them. We identify differences in predicted structure of RT in two
independent events of adaptation into the human hosts (HIV1/SIVcpz and HIV2/SIVsm) but none were common in between the two comparisons. The SInS were not evenly distributed across the five domains of the reverse transcriptase protein. There were no changes identified in the finger domain between HIV1/SIVcpz and between HIV2/SIVsm RTs suggesting that the finger domain of HIV1 and HIV2 have not changed in the course of evolution in the human host as compared to their ancestral counterparts. The predicted structural properties that largely differ at SInS were hydrophobicity and beta sheet. Changes in these secondary structural properties affect packing of protein. Beta sheets usually are involved in forming core structures of proteins. A more hydrophobic core structure with higher prediction of beta sheets suggest tight packing in proteins. We find differences in hydrophobicity and beta sheet in both cases of evolution and adaptation in humans (Table 5-3). This suggests very different packing in the two RT proteins, which in turn could be responsible for the different dimer stabilities that have been measured. Two pairs of SInS identified between HIV1/SIVcpz RT also showed proximity on the HIV1 RT crystal suggesting possible co-evolution of these sites affecting changes in the predicted secondary structures.

We also studied differences in RT between the two human immunodeficiency viruses (HIV1/HIV2) and the two simian immunodeficiency viruses (SIVcpz/SIVsm). The 19 SInS identified between HIV1/HIV2 RT shows predicted structural differences that have arisen in the course of independent evolution of RT in the two human immunodeficiency viruses. As 8 of the 19 identified SInS overlap with differences between SIVcpz/SIVsm RT, the remaining 11 SInS identified were the differences that
arose between the two human immunodeficiency viruses in the course of evolution in the human host. A few sites (3) out of the 11 sites clustered at the base of the thumb (Fig 5-2B). Such changes in domains interacting with the template can possibly explain the differences in fidelity rate between the two HIV RTs. The 11 sites were all present in the polymerase domain suggesting difference between the two HIV RTs were only in the polymerases domain (finger, palm, thumb, and connection) whereas the RNAseH domain was similar between the two based on the predicted secondary structure properties. The three sites (308, 313, 315) that cluster (distance < 7Å) at the base of thumb are interesting as structural changes at the base of the thumb domain between the HIV1/HIV2 RT could possibly address the functional differences (fidelity) between the two RTs.

We used predicted quaternary structures to test whether such instabilities at the dimer interface could be identified. Using 10 predicted properties of RT dimer we could separate cleanly the two different lineages (HIV1; SIVcpz and HIV2; SIVsm) but no difference could be observed within the lineages (Fig 5-3). The differences observed were largely due to difference in the number of atoms/residues between the two lineages. This would suggest that the differences in dimer strength measured in RT could be part due to differences in the number of atoms/residues interacting and forming bonds at the dimer interface between the two lineages. A lot of work still needs to be done to study how changes in secondary structure (SInS) can further inform us to changes in tertiary and possible quaternary structures.
Chapter 6

6.1 Summary

The goal of this thesis was to develop a statistical pipeline to identify whether evolution of viruses in different hosts is associated with changes in predicted structural components of their proteins. Identifying structural changes between phylogenetically related groups of proteins is difficult in the absence of crystal structure data. The situation is further complicated as no one to one relationship exists between change in sequence and change in structure [5,6].

We developed a statistical pipeline that, based on a LDA (Linear Discriminant Analysis) type technique, identifies differences in predicted secondary structure properties between phylogenetically related groups of proteins. The amino acid positions identified as structurally different in the predicted properties are called structurally informative sites (SInS). To develop the statistical pipeline we compared two subdomains of protein G of *Streptococcus* that had solved crystal structures [20]. We were able to identify as SInS the sites that are known to differentiate between the two structures of the subdomains. We tested the effect of different secondary structure prediction methods (servers) on our pipeline. We found that servers employing different methods show differences in their output sufficient to change the SInS identified by our pipeline from the same query sequence sets. Some servers outperform others in the query that we use (sub-domains of protein G), but the key takeaway point is that our pipeline is strongly affected by the choice of the prediction server. We also tested a range of within and between group sequence variability and its effect on prediction of SInS. We found
that for the range of variability tested in the study, there is no strong effect on prediction of SInS. We also found that, although we did not have false positives, we were missing some true SInS because of the stringent thresholds used in our pipeline. Expanding our approach to consider the CV (Coefficient of Variation), we were able to identify all true SInS between the two sub-domains of protein G.

The first viral dataset explored using this approach was that comprising proteins of the avian and human metapneumoviruses in chapter 3. Using this data we are able to identify SInS between the two groups. We were able to show that no strong correlation existed between sequence change and structure change as protein that were highly polymorphic in sequence showed similar number of SInS when compared to conserved proteins in the metapneumovirus genome. The predicted structural properties that changed frequently between the two groups of viruses affected the packing of the protein. The SInS also maintained signatures of phylogenetic group separation, even being able to separate out know subgroups within the human group. Moreover, SInS appeared to arise early in the course of adaptation, as they accumulated synonymous substitution at the same rate as other sites.

Finally we studied the structural evolution of the RT (Reverse Transcriptase) molecule in the primate lentivirus groups. We considered specifically two independent adaptation events, HIV1 from SIVcpz and HIV2 from SIVsm [108]. Comparing the structures of HIV1 and HIV2 we found evidence of differences in structure between the two that arose in the course of the two independent adaptation events. Some of these structural differences are maintained from differences between their respective ancestral
hosts. The predicted structural properties that differ at these SInS affect the packing of
the RT molecules in the primate lentivirus groups. Based on quaternary properties of the
protein we could separate out the protein into two groups based on their independent
lineages (HIV1, SIVcpz and HIV2, SIVsm).

In conclusion, the pipeline developed in this thesis helps identify structural
differences in phylogenetically related proteins starting only from available sequence
data. This is a novel tool for analysis of sequence which, along with phylogenetic
analysis, provides valuable information about early structural adaptive processes that
occur in the course of evolution in the absence of crystal structures.

6.2 Future Work

In the analysis of RT from primate lentiviruses a major question was how SInS
identified from secondary structure properties could inform us about changes in higher
order structures. As the crystal structure of RT is solved, we would like to investigate
what structural elements of the SInS affect the three dimensional tertiary structure. Our
results suggest that the differences in quaternary structure at the dimer interface are in the
numbers of contact residues. We plan to screen the position of SInS and check whether
they cause any differences in the contact at the dimer interface. This would help us
understand information about changes in higher order structures gleaned from SInS
identified based on differences in secondary structures. The pipeline also needs to be
automated, so with user defined query sequences we can directly generate SInS as an
output. This would help us analyze larger sequence datasets, and provide a user friendly
tool for the community.
Reference List


100 


VITA Sunando Roy

EDUCATION

The Pennsylvania State University, University Park, PA.
Ph.D. in Bioinformatics and Genomics. [GPA 3.88/4.00]
Dissertation: Evaluating structural evolution of proteins using multivariate statistics. (Advisor: Dr Mary Poss)

Tata Institute of Fundamental Research, Bombay, India.
Master of Science in Biology
Dissertation: Effect of optical and fluid forces on uninfected and malaria infected red blood cells. (Advisor: Dr Shobhona Sharma)

Abasaheb Garware College, Pune, India.
Bachelor of Science in Microbiology

AWARDS


Recipient of Training grant (2010) - USDA-AFRI Microbial Functional Genomics program (Award # 2010-65110-20488)

SELECTED PUBLICATIONS


SELECTED PRESENTATIONS
