Fall 12-22-2010

Modeling Virus-Host Networks

James P. Evans
University of Pennsylvania, evansjp@mail.med.upenn.edu

Modeling Virus-Host Networks

Abstract
Virus-host interactions are being cataloged at an increasing rate using protein interaction assays and small interfering RNA screens for host factors necessary for infection. These interactions can be viewed as a network, where genes or proteins are nodes, and edges correspond to associations between them. Virus-host interac- tion networks will eventually support the study and treatment of infection, but first require more data and better analysis techniques. This dissertation targets these goals with three aims. The first aim tackles the lack of data by providing a method for the computational prediction of virus-host protein interactions. We show that HIV-human protein interactions can be predicted using documented human peptide motifs found to be conserved on HIV proteins from different subtypes. We find that human proteins predicted to bind to HIV proteins are enriched in both documented HIV targeted proteins and pathways known to be utilized by HIV. The second aim seeks to improve peptide motif annotation on virus proteins, starting with the dock- ing site for protein kinases ERK1 and ERK2, which phosphorylate HIV proteins during infection. We find that the docking site motif, in spite of being suggestive of phosphorylation, is not present on all HIV subtypes for some HIV proteins, and we provide evidence that two variations of the docking site motif could explain phosphorylation. In the third aim, we analyze virus-host networks and build on the observation that viruses target host hub proteins. We show that of the two hub types, date and party, HIV and influenza virus proteins prefer to interact with the latter. The methods presented here for prediction and motif refinement, as well as the analysis of virus targeted hubs, provide a useful set of tools and hypotheses for the study of virus-host interactions.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Lyle Ungar

Keywords
HIV, network

Subject Categories
Bioinformatics | Computational Biology

MODELING VIRUS-HOST NETWORKS

James P. Evans

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2010

Supervisor of Dissertation

_____

Lyle Ungar, Associate Professor, CIS

Graduate Group Chairperson

_____

Maja Bucan, Professor, Genetics

Dissertation Committee

Frederic Bushman, Professor of Microbiology

Sridhar Hannenhalli, Associate Professor of Genetics

Sara Cherry, Assistant Professor of Microbiology

Aydin Tozeren, Professor of Biomedical Engineering, Drexel University

Acknowledgements

I thank my mentors, Lyle and Aydin, for their support and guidance during my graduate studies. Lyle taught me to focus on the main contributions of my work and the work of others, and how to choose the right null hypothesis. Aydin provided me with a deeper understanding of peer review, grant writing and reviewing, and initiating collaborations with other researchers. Both mentors helped me learn to translate projects into papers.

I thank my thesis committee members for their help with my dissertation and other projects. I'm glad I worked with Sridhar during my first rotation, and I'm thankful for his focus on controls. Rick and Sara's focus on the biological importance of computational projects will guide me during the rest of my research career.

I thank the students at Penn and Drexel for their friendship and research discussions. I enjoyed working with Greg while taking over his rotation project. I'm thankful that I was able to talk to Rithun about HIV. I'm very glad that Will has been around for scientific discussions.

I'm thankful that I found friends to accompany me during explorations of the east coast. I'm glad I met Logan, Rumen, Ryan, Kathleen, Mahdi, Yichuan, Noor, Adam, Mike, and Andrew. I'm glad Christin found me.

I thank my parents for their support during graduate school. They helped me look for an apartment when I started school. They traveled to Philadelphia when I had surgery, and stayed with me until I was back on my feet (literally). They came to see my defense despite my insistence that they did not need to.

I thank the Upenn hospital doctors for unraveling my intestines and finally locating and removing my appendix. Extreme abdominal pain has forced me to visit three hospitals, and my Upenn hospital experience has been the most successful.

Finally, I thank my funding sources. This work was supported by National Institutes of Health (NIH) grant # 232240, NIH training grant T32 HG000046, and National Science Foundation grant # 235327.

ABSTRACT

MODELING VIRUS-HOST NETWORKS

James P. Evans

Lyle Ungar

Virus-host interactions are being cataloged at an increasing rate using protein interaction assays and small interfering RNA screens for host factors necessary for infection. These interactions can be viewed as a network, where genes or proteins are nodes, and edges correspond to associations between them. Virus-host interaction networks will eventually support the study and treatment of infection, but first require more data and better analysis techniques. This dissertation targets these goals with three aims. The first aim tackles the lack of data by providing a method for the computational prediction of virus-host protein interactions. We show that HIV-human protein interactions can be predicted using documented human peptide motifs found to be conserved on HIV proteins from different subtypes. We find that human proteins predicted to bind to HIV proteins are enriched in both documented HIV targeted proteins and pathways known to be utilized by HIV. The second aim seeks to improve peptide motif annotation on virus proteins, starting with the docking site for protein kinases ERK1 and ERK2, which phosphorylate HIV proteins during infection. We find that the docking site motif, in spite of being suggestive of phosphorylation, is not present on all HIV subtypes for some HIV proteins, and we provide evidence that two variations of the docking site motif could explain phosphorylation. In the third aim, we analyze virus-host networks and build on the observation that viruses target host hub proteins. We show that of the two hub types, date and party, HIV and influenza virus proteins prefer to interact with the latter. The methods presented here for prediction and motif refinement, as well as the analysis of virus targeted hubs, provide a useful set of tools and hypotheses for the study of virus-host interactions.

Contents

Acknowledgements **ii**

Acronyms **xii**

1 Introduction **1**

1.1 Motivation 1

1.2 Organization 2

1.3 Review of virus-host interactions 5

1.3.1 Datasets 7

1.4 Previous work with virus-host interactions 10

1.4.1 Modeling virus-host interactions 14

2 Prediction of HIV virus-host protein interactions using virus and host sequence motifs **16**

2.1 Background 16

2.1.1 Virus-host network use cases 17

2.1.2 Reasons for predicting virus-host interactions 18

2.1.3 Predicting virus-host interactions 20

2.2 Methods 23

2.2.1 Virus protein motif annotation and conservation 23

2.2.2 Human protein peptide motif and domain annotation 25

2.2.3 Predicting interactions between HIV and human proteins . . . 26

2.2.4 Validation using the NCBI HIV-Human Protein Interaction Database . . . 28
2.3 Results . . . 31
2.3.1 Human peptide motifs were conserved on HIV proteins . . . 31
2.3.2 Significant overlap of predicted and validated HIV-human interactions verifies our method . . . 33
2.3.3 Predicted and validated HIV-human interactions share similar Gene Ontology labels . . . 35
2.3.4 Predicted and validated HIV-human interactions occupy the same KEGG pathways . . . 37
2.3.5 Virus level comparison of predicted and verified HIV-human interactions . . . 43
2.3.6 Infrequent host peptide motifs did not improve prediction performance . . . 47
2.4 Discussion . . . 48
2.5 Conclusion . . . 52

3 A bioinformatics approach reveals possible MAPK docking motifs on HIV proteins 55

3.1 Background . . . 55
3.1.1 MAPK and HIV infection . . . 56
3.1.2 MAPK substrate docking . . . 56
3.1.3 Disrupting protein-protein interactions using small-molecule inhibitors . . . 57
3.2 Results . . . 59
3.2.1 Consensus MAPK docking sites on human proteins . . . 59
3.2.2 MAPK docking sites on HIV proteins . . . 60
3.2.3 Candidate docking motifs on HIV Nef . . . 68

3.2.4 Candidate docking motifs on the HIV protein matrix are supported by structures . . . 68
3.3 Discussion . . . 70
3.4 Conclusion . . . 73
3.5 Methods . . . 74
3.5.1 Human and HIV sequences and motifs . . . 74
3.5.2 Significance of proposed docking site motif conservation on HIV proteins . . . 74
3.5.3 Structure analysis . . . 75
3.5.4 Docking . . . 76

4 Modularity in protein interaction network hubs predicts viral host-pathogen interactions 77
4.1 Background . . . 77
4.1.1 Intermodular and intramodular hubs . . . 78
4.2 Results . . . 81
4.2.1 Hub classification . . . 81
4.2.2 Hub class properties . . . 84
4.2.3 Virus hub preference . . . 84
4.3 Discussion . . . 89
4.4 Conclusion . . . 91
4.5 Methods . . . 92
4.5.1 Human interaction networks . . . 92
4.5.2 Virus-host interaction networks . . . 93
4.5.3 Peptide motif and SMART domain annotations . . . 94

5 Reflections and perspectives 95
5.1 Review of our work . . . 96
5.2 Future work . . . 99

A Supplemental tables **104**

List of Tables

1.1 Experimentally determined virus-host interactions 6

2.1 Overlap between predicted and validated HIV-human interactions . . 34
2.2 Gene Ontology molecular function enrichment for predicted and experimentally verified HIV targeted human proteins 36
2.3 Validation of HIV-human predicted interactions using Gene Ontology biological process similarity 37
2.4 KEGG pathway enrichment for predicted and experimentally verified HIV targeted human proteins 38

3.1 MAPK docking pattern hits on human proteins 59
3.2 MAPK docking pattern hits on HIV proteins 61

4.1 Fisher's test for virus hub preference 85
4.2 Fisher's test for virus hub dependency factor preference 88

A.1 HIV protein sequence counts 105
A.2 Counter domain/peptide motif relations and coverage 106
A.3 Validation of direct HIV-human interaction predictions 107
A.4 HIV-human interaction prediction validation with KEGG pathways . 108

List of Figures

1.1 HIV genome 7
1.2 HCV genome 9
1.3 Host peptide motifs on HIV NEF 12

2.1 Network diagrams for HIV-human protein interactions 27
2.2 Host peptide motif conservation on HIV NEF 31
2.3 Host peptide motifs conserved on HIV proteins 32
2.4 Evaluation of direct HIV-human interaction predictions 35
2.5 HIV TAT natural killer cell mediated cytotoxicity 40
2.6 HIV TAT T cell receptor signaling pathway 41
2.7 Comparison of predicted and validated virus-host interactions for host proteins in KEGG pathways 44
2.8 Comparison of combined predicted and validated virus-host interactions 45
2.9 Evaluating the use of infrequent host peptide motifs for HIV-human interaction prediction 46

3.1 MAPK docking site pattern hits on HIV proteins 62
3.2 Docking between MAPK ERK1 and HIV Nef 64
3.3 Human sequence logos for MAPK docking site hits 66
3.4 HIV sequence logos for MAPK docking site hits 67
3.5 MAPK substrate hierarchy 69
3.6 Docking between MAPK ERK1 and HIV MA 71

4.1 Comparing hubs across human interaction networks 81
4.2 Establishing inter/intramodular hubs 83
4.3 Virus hub preference 86
4.4 Virus-host hub network 90

Acronyms

CD Counter domain

DAVID Database for annotation, visualization, and integrated discovery

DHHE Direct HIV-human experimental interactions

dPCC Distribution of Pearson correlation coefficients

EBV Epstein-Barr virus

ELM Eukaryotic linear motif

GO Gene ontology

H1 Human proteins that directly interact with a virus protein

H2 Human proteins that interact with virus targeted human proteins

HCV Hepatitis C virus

HHE HIV-human experimental interactions

HHP HIV-human predicted interactions

HIV Human immunodeficiency virus

HPRD Human protein reference database

I2D Interologous interaction database

KEGG Kyoto encyclopedia of genes and genomes

LANL Los Alamos national laboratory

MAPK Mitogen-activated protein kinase

NCBI National center for biotechnology information

PCC Pearson correlation coefficient

PPI Protein-protein interaction

siRNA Small interfering RNA

STRING Search tool for the retrieval of interacting genes/proteins

VDF Virus dependency factor

Chapter 1

Introduction

1.1 Motivation

Virus proteins interact with their host's cellular machinery, and in doing so alter the host cell to favor viral replication. It is important to identify and annotate virus-host interactions for the discovery of new drug targets as well as for assessing the efficiency of antiviral drug therapies on host subpopulations [19, 37]. Furthermore, virus-host interactions can be used to suggest roles for virus proteins [26], and investigate common viral strategies for interacting with their hosts [116]. With this in mind, many researchers have gathered data for virus-human interactions to determine host factors necessary for the virus life cycle. These interactions take the form of virus-human protein interactions, human gene expression responses to infection, and genetic screens for human genes that are necessary for virus survival. The genetic screens suggest that certain human genes play a role in infection, while the protein interaction and gene expression data suggest what these roles might be. To date, only three viruses, human immunodeficiency virus (HIV), hepatitis C virus (HCV), and influenza A virus have extensive data for these interactions.

The experimental challenges of identifying virus-host protein interactions are numerous. The biggest challenge is designing screens stringent enough to have low false

positive rates while ensuring that the number of real interactions that cannot pass these stringent assays is kept to a minimum. Protein interactions are transient in nature, and moreover such protein binding interactions may depend on the presence of cofactors that are not necessarily present in binding assays. These difficulties suggest supplementing experiments with a bioinformatic approach to predicting and understanding virus-host interactions. This has been attempted for HIV by predicting HIV-human protein interactions, but it was found that the most predictive human protein feature was the number of interacting partners a protein had [160], which does not provide sites for drug targeting. Such findings indicate that more analysis is needed to understand the underlying principles of interactions, and to guide experimental studies.

The overall goal of this dissertation is to estimate and analyze virus-host interactions. These interactions have been organized into networks, where proteins are nodes and network edges represent interactions between proteins [67]. We begin this thesis by predicting virus-host protein interactions using short peptide motifs that have been shown to play a role in protein interactions [82]. We demonstrate the validity of our predictions using HIV-human interactions. Next we investigate the usage of mitogen-activated protein kinase docking motifs on HIV proteins. We then compare HCV, HIV, and influenza virus-host networks to address the finding that virus proteins prefer to interact with highly connected host proteins. We provide evidence that this hub preference is a consequence of the requirement of viruses to utilize host protein complexes during infection. We conclude this dissertation with a discussion of how the work presented here aids virus-host network analysis.

1.2 Organization

This thesis has been organized into three specific aims: (1) the prediction of virus-host interactions using sequence motifs, (2) the refinement of virus sequence motifs,

and (3) the analysis of virus targeted host proteins.

This chapter serves as a background section. We will discuss two types of virus-host interactions: protein interactions gathered from high throughput studies and literature searches, and genetic screens searching for host factors required for the viral life cycle. Next, we will cover previous studies of virus-host networks, describing network and gene analysis and the importance of peptide motifs in virus-host protein interactions. We will then conclude with a description of methods for predicting protein interactions, and illustrate how they have been applied to virus-host interactions.

Chapter 2 is devoted to the first aim, the prediction of virus-host protein interactions, which are important for guiding experimental studies [79, 96]. We focus this aim on interactions between HIV and human proteins and show that a list of host proteins highly enriched with those targeted by HIV proteins can be obtained by searching for host protein motifs along virus protein sequences. We find that peptide motifs conserved across 70% of HIV protein sequence samples occur in similar positions on HIV proteins, and we document protein domains that interact with these conserved motifs. We predict which human proteins may be targeted by HIV by taking pairs of human proteins that may interact via a peptide motif conserved in HIV and the corresponding interacting protein domain. Our predictions are enriched with host proteins known to interact with HIV proteins Env, Nef, and Tat (p-value $<$ 4.26e-21). Cellular pathways statistically enriched for our predictions include the T cell receptor signaling, natural killer cell mediated cytotoxicity, cell cycle, and apoptosis pathways. Molecular functions enriched with both predicted and confirmed HIV targeted proteins include phosphorylation and adenyl ribonucleotide binding. This study validates the role of peptide binding motifs in guiding virus-host interactions and suggests new HIV targeted pathways and proteins in the host cell.

In Chapter 3 we discuss the second aim, the refinement of virus motifs. Over the

course of HIV infection, virus replication is facilitated by the phosphorylation of HIV proteins by human ERK1 and ERK2 mitogen-activated protein kinases (MAPKs). MAPKs are known to phosphorylate their substrates by first binding with them at a consensus docking site motif. Docking site interactions could be viable drug targets because the sequences guiding them are more specific than phosphorylation consensus sites. In this study we use multiple bioinformatics tools to discover candidate MAPK docking site motifs on HIV proteins known to be phosphorylated by MAPKs, and we discuss the possibility of targeting docking sites with drugs. Using alignments of multiple HIV protein sequences taken from different patients, we show that the consensus MAPK docking pattern previously described for human proteins is missing from a significant fraction of the sequences gathered for HIV proteins known to be phosphorylated by ERK1 and ERK2. We revise the consensus MAPK docking pattern in order to provide patterns that annotate that the majority of sequences for all HIV proteins. One revision is based on a documented human variant of the consensus MAPK docking motif, and the other reduces the number of required basic amino acids in the consensus docking motif from two to one. The proposed patterns are shown to be consistent with *in silico* docking between ERK1 and the HIV matrix protein. The motif usage on HIV proteins is sufficiently different from human proteins in amino acid sequence similarity to allow for HIV specific targeting using small-molecule drugs.

We tackle the third aim, the analysis of virus targeted highly connected human proteins, in Chapter 4. HIV, influenza virus, and other human viruses preferentially interact with highly connected proteins, or hubs, in the human protein interaction network [26, 40, 50, 160]. Hub proteins have been classified into two groups by co-expression with their interacting neighbors [162]. Intermodular, or date, hubs are defined as being co-expressed with their neighbors in certain tissues, while intramodular, or party, hubs are characterized by co-expression with their neighbors in most tissues [162]. The existence of these two hub classes is under

debate [1, 12, 13, 67, 162]. Here we provide new evidence for this hub distinction, and ask if hub proteins that interact with virus proteins have a significant association with one of the hub classes. For HIV and influenza virus, we show that hubs that directly interact with a virus protein are more likely to be intramodular than intermodular. This preference is important because the features of intramodular hubs that are different from intermodular hubs serve as hypotheses for features that virus proteins target in host proteins. The virus intramodular hub preference is also important for the study of biological networks because it is further evidence for the existence of two hub classes.

The thesis concludes in Chapter 5 with a review of the material presented and a discussion of future work for virus-host interactions. We will summarize our major discoveries and testable hypothesis generated in Chapters 2, 3, and 4. Next, we will discuss future work that stems from this dissertation. We will show how protein structure and virus-host network patterns can be used jointly with the peptide motifs from the first aim to arrive at a new set of virus-host interaction predictions. We will cover how peptide motifs on virus proteins can be validated by constructing a database of virus protein mutations and their effects on virus-host interactions, and we will discuss a project to test the hypothesis that virus proteins serve as intermodular hubs in virus-host networks. We will also propose a project to test that hypothesis that influenza virus uses different versions of peptide motifs based on which host it infects.

1.3 Review of virus-host interactions

While single species protein-protein interaction networks have been gathered in high throughput screens and studied for ten years, virus-host networks are only recently the subject of investigation [111]. Large scale studies of virus-host interactions have focused on HIV, HCV, and the influenza virus. Smaller studies have been conducted

Virus	siRNA screen hits	Protein interactions (Interacting human proteins)
HIV	850	2652 (887)
HCV	318	477 (414)
Influenza	295	339 (230)
Papillomavirus	NA	229 (94)
EBV	NA	173
KSHV	NA	173
VZV	NA	123

Table 1.1: For each virus, we show the number interactions that occur between virus and human proteins, and count the number of host factors necessary for infection (siRNA screen hits). Of the viruses in the table, HCV, HIV, and influenza virus have the most experimentally determined virus-host interactions. siRNA screen hits describe interactions between a virus and a human gene. siRNA screens have only been conducted for HCV, HIV, and the influenza virus. Protein interactions cover only direct interactions, like protein binding or protein modifications, such as the phosphorylation of a virus protein by a host kinase. A single human protein can be involved in multiple interactions with different virus proteins, so the total number of unique human proteins involved in virus-human interactions is given in parentheses. While only direct interactions are listed here, HIV has a total of 3950 direct and indirect (e.g. regulatory, induced protein modification) interactions with 1439 human proteins [62].

for Epstein-Barr virus (EBV) [26], Kaposi sarcoma-associated herpesvirus (KSHV), Varicella-Zoster virus (VZV) [171], and Papillomavirus [47]. Virus-host interaction data have primarily been collected from small interfering RNA (siRNA) screens, high throughput binding assays, and literature reviews. siRNA screens were performed in infected cells to find host factors that could be knocked down without harmful effects on the host, but would inhibit virus replication [63]. Table 1.1 enumerates virus-host interactions gathered for different viruses.

We focus on HIV, HCV, and influenza virus for two reasons. First, unlike EBV, VZV, and KSHV, all three viruses have small proteomes, making them more reliant on host cell machinery during their life cycles. Second, the interactions of HCV, HIV, and influenza virus with human proteins are well studied, and include both protein interactions and siRNA data. The following section describes the interactions for these three viruses in detail, and later we discuss principal findings from previous studies of these networks, and initial attempts to model them.

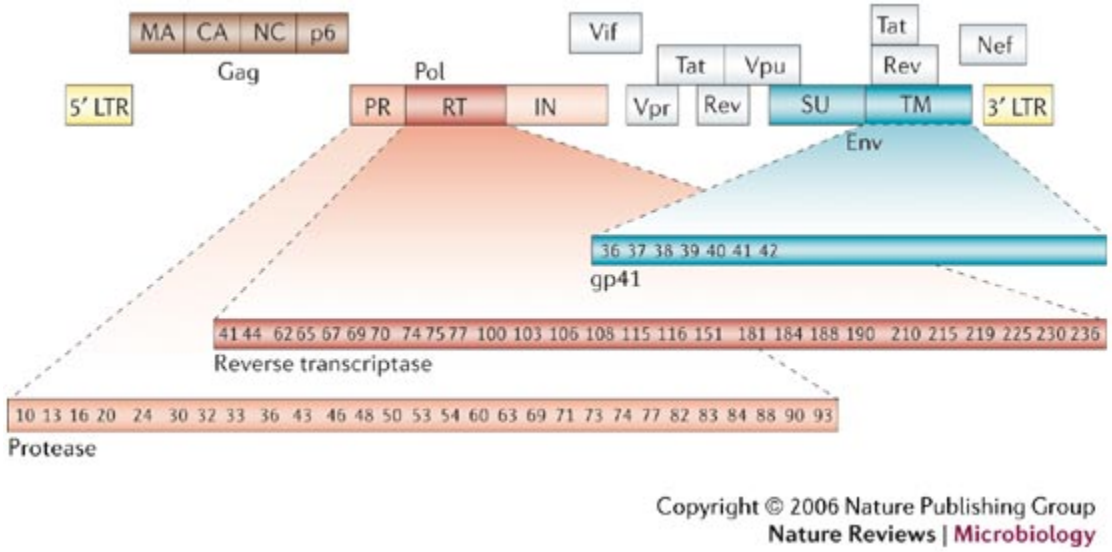


Figure 1.1: HIV contains nine open reading frames and produces around twenty proteins. This image was taken from an HIV review article [98]. Permission to reuse this figure was granted by the publisher, copyright 2006 Nature Publishing Group.

1.3.1 Datasets

HIV-human interactions

HIV is an RNA virus of roughly nine kilobases encoding nine open reading frames that produce around twenty proteins (Figure 1.1) [58]. There are different subtypes, or strains, of HIV, and these are classified hierarchically, starting with three groups: major (M), outlier (O), and non-major and non-outlier (N) [161]. Group M, which is the most common, has been divided into nine subtypes, or clades: A, B, C, D, F, G, H, J, and K. Sequences from the same subtype are more similar to each other than to sequences in other subtypes. Some subtypes correspond to geographical locations. Recombinant forms of group M subtypes have been identified. For instance, 01_AE is a combination of subtypes A and E that is circulating in Southeast Asia [161]. Subtype E has not been found in a non recombinant form, so it is not listed in the nine subtypes of group M. Five subtypes and two recombinants are present in at

least 2.5% of the world population [161].

Interactions between human proteins and HIV proteins come mostly from literature curation rather than high throughput binding assays. Interactions between HIV and human proteins have been cataloged in VirusMINT [29] the NCBI HIV-Human Protein Interaction Database [130], and the pathogen interaction gateway (PIG) [47]. From these databases, it has been observed that HIV proteins interact with many of the same host proteins [62]. HIV-human interactions come in two types, direct and indirect. Direct interactions involve physical protein contact, and include binding interactions, protein modifications, and cleavage interactions. Indirect interactions involve gene expression regulation and indirect effects, such as inducing protein modifications or cleavage. In HIV-human interaction databases, the number of human proteins involved in indirect virus-host interactions is roughly twice the number that are involved in direct interactions [62].

Not all of these interactions will occur *in vivo*, or be relevant to HIV infection. Direct interactions pertinent to infection can be found by comparing these protein interaction databases to siRNA screens for host factors involved in HIV replication. Four large such screens, each resulting in around two hundred host factors have been conducted for HIV [19, 92, 181, 184]. There was little overlap between the four screens [24, 181]. While this might be attributed to differences between the cell types used, more than 90% of the genes deemed important for HIV infection were expressed in all cell types [24]. Other explanations for different results include experimental error, different filtering thresholds for deciding which host genes resulted in cell lethality when knocked down, and differences in the infection time points analyzed [24].

HCV-human interactions

HCV is a 9.6 kb positive-strand RNA virus that encodes a 3000 residue polyprotein, which is cleaved by host and virus proteases into ten proteins (Figure 1.2) [113].

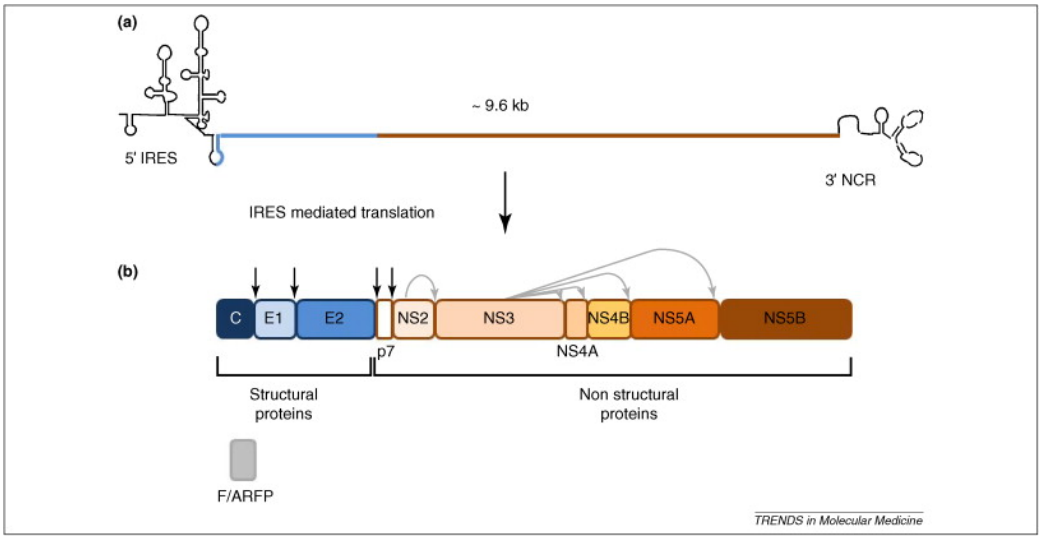


Figure 1.2: (a) HCV genome and (b) HCV polyprotein. This image was taken from a review of HCV-human interactions [63]. Permission to reuse this figure was granted by the publisher, copyright 2010 Elsevier Ltd.

Chronic HCV infection leads to serious liver disease [63], with roughly three percent of the world's population infected [151]. In the last decade, several infectious model systems have enabled the accumulation of HCV-human interactions [63].

The HCV-human protein interaction network has been constructed from high-throughput screens augmented with a literature curation [40]. Unlike HIV, there is little overlap between the host binding partners of HCV proteins, and all HCV-human interactions are for direct protein binding. Several small siRNA screens have been combined with a larger genome-wide screen to arrive at over three hundred host factors required for HCV infection [102]. While the HIV genetic screens focused on host dependency factors, some HCV siRNA screens looked for host factors that when knocked down caused an increase in virus replication. These were likely host genes that were part of the immune response. One such screen identified twenty five immune response genes [102].

Influenza-human interactions

Influenza A virus is a negative-strand RNA virus that encodes eleven proteins using eight individual RNA segments [31]. An early influenza virus siRNA screen, conducted in fly rather than human cells, identified 100 fly genes involved in influenza virus replication [70]. Two recent studies have tried to pin down the human pathways and genes involved in influenza A infection. A genome-wide siRNA screen in cells infected with influenza A virus identified 295 host factors required for influenza replication [91]. A more detailed study extensively cataloged three types of interactions between host and virus: protein-protein, host gene expression response, and siRNA screens [146]. Like the HCV siRNA data, this study looked at both positive and negative virus response to host gene knock down, providing lists of necessary host factors and possible immune response genes. Analysis of the data revealed many aspects of the host immune system to interact with the virus.

1.4 Previous work with virus-host interactions

It is important to study and model virus-host interaction networks at protein, the pathway, and network levels to understand virus protein function [26], help design antiviral therapies [19, 37], guide virus-host protein interaction experiments [79, 96], and compare the ways in which viruses alter host cellular pathways [116]. Each interaction abstraction has yielded important insights. Protein level studies often describe binding sites on both virus and host proteins. Pathway level studies have revealed subsets of human proteins that are likely to interact with virus proteins. These studies can be used as a guide for more detailed experiments at the protein level, and to compare virus-host protein interactions in terms of biological processes instead of individual proteins. Network level studies have allowed viruses to be compared to find trends common to infection.

Protein studies

At the protein level, studying the binding regions on virus and human proteins will aid in finding small-molecule drugs to prevent virus-host interactions. A recent study produced a number of U.S. Food and Drug Administration approved small-molecule drugs that could inhibit certain protein interactions after screening these drugs for their ability to disrupt protein complexes with known peptide binding sites and available 3D structures [126]. Without knowledge of the protein binding sites, this study would not have been possible.

The bulk of HIV protein interactions has come from collections of single protein studies [29, 47, 111, 130]. In addition to helping to build the HIV-human protein interaction network, these individual interaction studies have given insights into how virus-host protein interactions occur. It is now proposed that binding between host and virus proteins occurs between short peptide motifs on virus proteins, and protein domains on human proteins [82, 149, 166]. Figure 1.3 shows selected motifs on the HIV NEF protein. Each motif is associated with a domain or set of proteins that interact with it. A database of host peptide motifs and the domains that interact with them exists at the Eukaryotic Linear Motif (ELM) Resource [131]. The ELM Resource has cataloged over 130 of these peptide motifs and constructed a pattern that matches each one using documented motif instances from the literature. Work with binding regions on virus and host proteins has been hampered by a disconnect between the eukaryotic work done with peptide motifs and the study of virus-host interactions. In this dissertation, we combine knowledge from the ELM Resource with virus-host protein interactions to study the ability of peptide motifs to explain interactions between virus and host proteins.

Pathway studies

Knowledge of which host cellular pathways are targeted by viruses helps to narrow the focus of experiments determining virus-host protein-protein interactions [96].

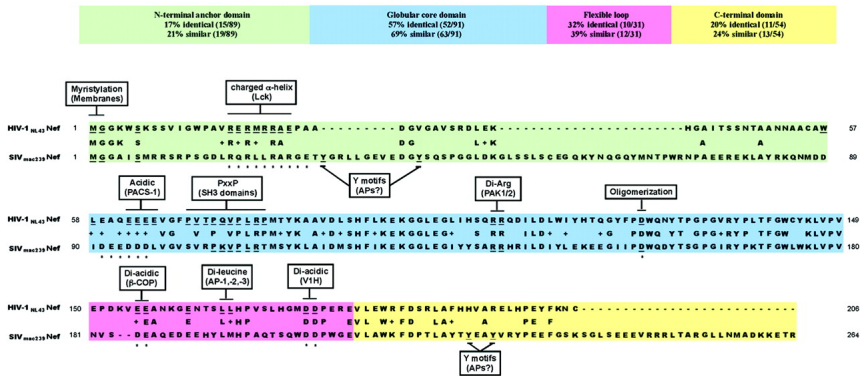


Figure 1.3: HIV and SIV NEF harbor short peptide motifs that enable them to interact with host cell proteins and alter host trafficking. This image was taken from a review of the effects of NEF on host cell trafficking [140]. Permission to reuse this figure was granted by the publisher, copyright 2006 American Society for Microbiology.

Experimental methods for determining protein interactions are costly and require much time and effort, so methods to guide experiments are desirable [153]. Mapping the yeast interaction network required several high throughput yeast two-hybrid screens due to the false negative rates of these assays [33, 76]. Focusing on specific pathways important for infection will reduce the cost and time required for such experiments.

Pathways are also useful for comparing viruses, or multiple siRNA screens investigating host factors required by a virus. Comparing HIV siRNA screens at the gene level showed little overlap between results, but at the pathway level results became more consistent [181]. Some common HIV targeted pathways included NFκB signaling, estrogen-receptor signaling, peroxisome proliferator-activated receptor signaling, RAR activation, and caspase apoptosis routes [181]. Similarly, pathway analysis of siRNA screens in HCV and other Flaviviridae viruses identified TGFβ, ErbB, MAPK, focal adhesion, and ubiquitin-mediated proteolysis as common Flaviviridae virus pathways [102]. Influenza virus, HIV, EBV, and KSHV, have also been found to target similar immune response pathways [18, 146].

Pathway studies have also facilitated the annotation of virus proteins of unknown

function [26], and the comparison of viral strategies for subverting the human type I interferon response [116]. Investigations of virus targeted pathways have led to hypotheses concerning virus protein function. Analyzing the pathways that were found to interact with Epstein-Barr virus (EBV) proteins allowed investigators to infer roles for unannotated EBV proteins [26]. Another pathway study of virus protein interactions with host proteins involved in the human type I interferon response revealed that virus proteins from four viral families (Flaviviridae, Herpesviridae, Papillomaviridae, and Retroviridae) targeted different aspects of the response, which consists of a four level cascade of interactions traveling from host receptors, to adapters and mediators, and ending at transcription factors that initiate the immune response [116]. Viruses in the Flaviviridae and Herpesviridae families targeted host proteins with many host interactions in the interferon pathway, focusing on protein adapters, mediators, and transcription factors. Viruses in the Retroviridae and Papillomaviridae families mostly targeted transcription factors.

One issue holding back pathway level studies is the availability of experimental data. High throughput yeast two-hybrid and siRNA screens have only been conducted for a handful of viruses. It would be nice to compare viruses without such datasets by only using sequence information, which is becoming easier to gather [115]. In this dissertation, we address this problem by describing a way to use peptide motifs on virus proteins to predict virus targeted pathways.

Network studies

While these pathway level studies are important, to understand infection as a system, researchers began investigating the virus-host interaction network. One network based analysis of virus-host networks searched for network motifs in the HIV-human interaction network [172]. Network motifs are statistically over-represented interaction patterns in networks. An example is a feed back loop in a regulatory network,

where a gene controls the expression of its regulator. Network motifs have been identified in yeast protein interaction networks [177], E. coli transcriptional regulation networks [150], and synthetic genetic interaction networks [35]. In an initial study of cross-species network motifs, the HIV-human network was searched for motifs that might aid the virus in taking control of the host cell.

Virus-host networks have also been used to achieve a more biological understanding of virus-host interactions. A combined study of HIV siRNA screens and the host protein interaction network clustered the virus targeted host network into dense network neighbor hoods. Analysis revealed these neighborhoods to represent proteasome, mediator, RNA binding and splicing, and chaperone network components [24]. With a similar goal in mind, an influenza virus-host network that included virus-host protein interactions, siRNA screen results, and host expression response to infection was analyzed to determine which interactions could be attributed to the host immune system [146].

In a study of the network properties of pathogen targeted proteins it was determined that pathogens like HIV and EBV have proteins that prefer to interact with human proteins with specific network properties. Hub proteins and bottleneck proteins, i.e. proteins that separate large components of the host interaction network, were found to be preferentially targeted by pathogen proteins [50]. A later topological analysis of the HCV-human interaction network revealed that HCV proteins also preferred to interact with human hub and bottleneck proteins [40]. It has been suggested that hubs are targeted by viruses because they provide an efficient way to rewire the host network to favor virus production [26]. In this dissertation, we address the virus hub preference, and suggest a biological reason behind it.

1.4.1 Modeling virus-host interactions

Just as single species protein interaction networks inspired methods to predict protein interactions, host-pathogen networks have initiated the search for network and

protein features that model and predict virus-host protein interactions. One study examined HIV-human interactions to determine what properties, or features, of human proteins made them more likely to interact with virus proteins [160]. The features examined included Gene Ontology labels [7], global gene expression profiles, human interaction partners, human protein domains, and HIV protein motifs, but it was determined that the most predictive feature was host protein degree, i.e. the number of host proteins that interact with a candidate host protein. This finding was consistent with other work done with viruses and host hub proteins, but had the unfortunate effect of predicting that all viruses will interact with the same host proteins. Another HIV-human protein interaction prediction study used structural similarity between host and virus proteins [46]. For each HIV protein with an available structure, the most structurally similar host proteins were found. The protein interaction neighbors of these host proteins were predicted to bind to HIV proteins. This method is limited to virus proteins with determined structures, and virus protein structures are hard to predict because so many of them are unstructured proteins [165], so a sequence based approach is preferable.

The work presented in this dissertation continues the exploration of virus-host interactions at the protein, pathway, and network level. First, we examine conserved host peptide motifs on HIV proteins. We show that these motifs can be used to predict HIV-human interactions, and we make an argument that some motifs should be refined. Then we seek an explanation for the observation that virus proteins target host hub proteins.

Chapter 2

Prediction of HIV virus-host protein interactions using virus and host sequence motifs

2.1 Background

An important component of systems biology is the determination and study of protein-protein interactions (PPIs) and the networks, or interactomes, that they form. Previous work with single organism systems has revealed PPI networks to be useful for annotating proteins of unknown function [86, 148], comparing organisms [48, 87, 147], predicting other interaction types [175], investigating the peptide regions guiding interactions [45, 105, 119], and identifying protein complexes [93]. The study of single organism networks has been extended to multiple organism host-pathogen networks, where virus and cellular parasite proteins alter host interaction networks by competing with host proteins for interactions in the host network [36, 154, 168]. As experimental work with virus-host PPI networks has grown, the methods for protein functional annotation and network comparison developed using

single organisms have been transferred to virus-host systems to generate hypotheses about virus protein function [26] and investigate common viral strategies for countering the host immune system [116].

2.1.1 Virus-host network use cases

The study of virus-host networks has not only aided antiviral drug discovery and treatment optimization using existing drugs [19], but furthered virology as well. Here we outline two examples where the analysis of virus-host PPI networks yielded new insights about viruses. The first case illustrates how virus-host networks can be used to form hypotheses about the functions of Epstein-Barr virus proteins. The second case describes a comparative study of viral mechanisms for dealing with the host immune system that was facilitated by knowledge of virus-host interactions.

Epstein-Barr virus (EBV), which has been linked to several diseases, including cancer, is a herpesvirus that has almost 90 proteins [26]. 43 of these proteins are conserved across most herpes viruses, and their functions have been investigated. However, the functions of the remaining 46 proteins are not as well understood [26]. Determining functions these proteins was made easier by generating and evaluating a virus-host network to help formulate testable hypothesis about virus protein function [26]. Using the EBV-human PPI network, some EBV proteins were hypothesized to have roles in cell survival and apoptosis because the human proteins they interacted with were known to have these functions [26]. The possible involvement of these EBV proteins in promoting cell survival and suppressing apoptosis is important because these activities may be aiding the progression of some cancers [182]. This transfer of human protein function to virus proteins using virus-host PPI networks has laid the ground work for future studies of virus proteins with unknown functions.

Virus-host PPI networks have also been utilized to compare viral strategies for subverting the human type I interferon response, which consists of a four level cascade of interactions traveling from host receptors, to adapters and mediators, and

ending at transcription factors that initiate the immune response [116]. Navratil et al. constructed a human type I interferon PPI network that included human proteins involved in the type I interferon response, as well as human proteins that interacted with these immune system proteins [116]. Then the interactions of proteins in this immune response network with virus proteins from four viral families (Flaviviridae, Herpesviridae, Papillomaviridae, and Retroviridae) were compared to find which of the four levels of the type I interferon response were targeted differently between viruses. All four virus groups were found to preferentially interact with the signaling part of the immune system. Viruses in the Flaviviridae and Herpesviridae families targeted host proteins with many host interactions in the interferon network, focusing on protein adapters, mediators, and transcription factors. Viruses in the Retroviridae and Papillomaviridae families mostly targeted transcription factors. As more virus interactions with host pathways are accumulated, this comparison of pathway specific virus-host interactions can be extend to other host cellular processes, such as apoptosis and autophagy.

2.1.2 Reasons for predicting virus-host interactions

For studies of virus-host networks to continue, and to ensure that the conclusions drawn from such networks are accurate, more interaction data are needed. HIV has nearly fifteen hundred host proteins that interact with its proteins [62, 130], but other viruses with similarly sized proteomes, like hepatitis C virus (HCV), have less than 500 virus-host interactions. This discrepancy in virus-host interactome size is partially caused by the excessive study of HIV-human interactions in comparison with other virus-host networks [111], but an additional contributor might be the way in which virus-host interactions have been collected for HIV. With the exception of HIV, virus-host interaction data have mostly been generated by yeast two-hybrid screens. For instance, a recent study investigating interactions between influenza and human proteins identified less than 350 PPIs using a stringent two-hybrid assay

that required interactions to be present in primary and secondary screens [146]. Such stringency in screening is required because of the significant false positive rates for high throughput screens like yeast two-hybrid and tandem affinity purification assays [72]. Stringent screens like the one used to identify human-influenza virus interactions have low false discovery rates, with estimates less than 15% for yeast and worm protein interactions [76], but they often rule out many true interactions, with false negative rates above 40% for yeast and worm [76]. False negative rates have been a problem for yeast and human networks, yielding incomplete networks that caused revisions of network properties and debates over conclusions as networks grew in size [1, 12, 13]. To make lasting conclusions from virus-host interactions, we need nearly complete virus-host interactomes. Due to the high false negative rates for high throughput screens, and based on the multiple screens required to construct a high quality yeast PPI network, several large-scale experiments will be required to accurately map a virus-host network [33], but making predictions for virus-host interactions can aid in accomplishing this goal by reducing cost and labor.

Experimental methods for determining protein interactions are costly and require much time and effort, so methods to guide experiments or replace them are desirable [153]. Predicted interactions for yeast have helped to improve the accuracy, coverage, and efficiency of PPI screens when used in combination with experiments [79, 96], and this will likely translate to virus-host networks. There are two ways in which PPI predictions can help in gathering additional virus-host interactions. First, predictions can serve as an additional validation of two-hybrid results. In recent high throughput influenza-human PPI assays, interactions were required to pass two screens [146]. This approach to PPI investigation has high false negative rates, but this could be solved by combining the screens with predicted interactions. Instead of discarding all primary screen interactions that failed to pass the secondary screen, only those primary interactions with no prediction support would be thrown out. These saved predicted interactions could then be tested in a third screen. PPI predictions can

also help to focus experiments that are more thorough than two-hybrid screens, such as luciferase complementation [112]. Computational approaches have helped by reducing the number of host proteins to verify experimentally [96]. Predictions could be used to find host pathways with which a virus interacts, and instead of using the whole host proteome in assays, only host proteins appearing in the prediction enriched pathways could be interrogated. In this study, we describe a new method for how such predictions can be made for virus-human interactions.

2.1.3 Predicting virus-host interactions

Previous host-pathogen interaction prediction methods focused largely on finding PPIs between human and cellular parasite proteins. One method found the probability that two protein domains interact given the human PPI network, and used this probability to find the likelihood that pathogen and human proteins interact given their domain profiles [49]. Another method used structures of human complexes as templates to match possible host-pathogen interactions against, under the hypothesis that a candidate host-pathogen interaction that resembles a host interaction is likely to represent a real host-pathogen interaction [38]. Candidate interactions consisting of a pathogen and host protein were matched against template host complexes using structural and sequence similarity. Candidate interactions that were similar to a template host complex were then subjected to a test that ensured that pathogen and host protein were both expressed in the same tissue and at the correct time in the pathogen's life-cycle. Translating these methods to interactions between virus and human proteins has been difficult because virus proteins have few domains and their structures are either unsolved, or hard to find by comparative modeling. For instance, to find structures for the N-terminal and C-terminal regions of HIV VIF, two different protein structures were required for comparative modeling [107]. Due to problems with missing domains and structures for virus proteins, in this chapter we address the utility of an interaction prediction method that examines sequences

instead of structures, and utilizes small virus peptide motifs that guide interactions instead of domains.

A study of HIV-human interactions conducted by Tastan et al. tried to find the most predictive feature of virus-host interactions [160]. In their method, each interacting virus-host protein pair was associated with a feature vector composed of parameters related to Gene Ontology (GO) [7], global gene expression profiles, the human interaction network, human protein domains, and HIV protein motifs. Using roughly one thousand direct HIV-human interactions taken from the NCBI HIV-Human Protein Interaction Database [62, 130] as a training set, they determined that the most predictive feature of virus-host interactions was the number of host proteins with which the virus-host protein pair's human protein interacted. The more interactions a human protein had in the human interaction network, the more likely it was to interact with a virus protein. While this is consistent with other results from work with hepatitis C virus, Epstein-Barr virus, and other human viruses that showed that virus targeted proteins had significantly more host interactions that other proteins [26, 40, 50], it is not very useful for comparative studies of virus-host interactions because it predicts that all viruses interact with the same host proteins. For features that differ between viruses, like virus peptide motifs that guide protein interactions [82, 149, 166], Tastan et al. estimated a relatively weak potential for predicting virus-host interactions. Here we reevaluate this finding by using not only the direct virus-host interactions evaluated by Tastan et al., but indirect, regulatory interactions as well.

We focus this chapter on the computational identification of host proteins targeted by an invading virus. We use HIV infection as a case study because extensive study at the molecular level has yielded nearly fifteen hundred HIV targeted human proteins, covering nearly four thousand experimentally determined HIV-human interactions, which are cataloged in the NCBI HIV-Human Protein Interaction Database.

We predict virus-host interactions based on PPIs mediated by short eukaryotic linear motifs (ELMs) [131] on HIV proteins and human protein counter domains (CDs) known to interact with these ELMs. It has been estimated that 15% to 40% of host protein interactions are mediated by interactions involving a peptide motif [27, 121, 128]. The ELM Resource has cataloged over 130 of these peptide motifs and constructed a pattern that matches each one using documented motif instances from the literature [131]. The ELM Resource has also documented the protein CDs that interact with each motif. We aim to obtain human protein sets enriched with sets of known virus targeted proteins by annotating ELMs on HIV proteins, and using CDs on human proteins to match them with HIV proteins based the ELM Resource's catalog of ELM and CD associations.

The potential functional roles of interactions mediated by ELMs and their CDs in viral infection have been addressed in a number of recent articles [82, 149, 166]. The HIV literature contains at least ten examples of HIV-human PPIs that are directly associated with motif and domain presence. The motif/domain basis of such PPIs is not restricted to a single HIV protein, but is widely distributed across the HIV proteome, including HIV NEF [140], ENV [25], TAT [169], REV [169], VIF [110], and VPU [55]. This experimental evidence is the motivation for systematically investigating the association of motif/domain pairs with PPIs between virus and host proteins. Although Tastan et al. [160] estimated a relatively weak link between binding motif presence and the actual virus-host PPIs, their work was restricted to predicting direct binding between host and HIV proteins. In this study, we set out to identify all virus-host interactions, both direct and indirect, that have been documented between HIV and human proteins. Our hypothesis-based approach requires no training data for virus-host interactions to predict interactions. We only need virus and host protein sequences and the host interactome. As such, it is directly applicable to identifying host protein sets enriched with virus targeted host proteins for a wide scope of infectious diseases. The extremely low p-values we

calculate for the overlap between our predictions and experimentally verified HIV-host protein interactions, and the statistically significant Gene Ontology similarity we find between our predictions and experimentally validated interactions indicate the potential value of our approach for guiding the experimental detection of virus-host interactions and understanding the protein regions involved in these interactions.

2.2 Methods

2.2.1 Virus protein motif annotation and conservation

As a first step in predicting HIV-human interactions using relations between peptide motifs on HIV proteins and domains on human proteins, we used 133 peptide motif patterns from the ELM Resource to annotate HIV protein sequences taken from multiple patients and found peptide motifs that were conserved across patients, assuming that such conservation would be indicative of function. For each of nine HIV open reading frames (ENV, GAG, NEF, POL, REV, VIF, VPR, TAT and VPU), we downloaded the 2007 versions of alignments of hundreds of protein sequences spanning multiple patients and years from the Los Alamos National Laboratory (LANL) HIV Sequence Database (http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html) and removed all sequences except those labeled as subtypes B or C. We focused on subtype B because it is most common in the industrialized world [75], and chose subtype C because it is most common globally [74]. The HIV GAG and POL polyproteins are cleaved by proteases to produce smaller proteins [58]. GAG is cleaved into proteins CA, MA, NC, P1, P2, and P6, while POL cleavage produces proteins IN, PR, and RT. The LANL database provides alignments of uncleaved GAG and POL proteins, so to construct the full HIV proteome, we computationally split the GAG and POL alignments into their respective cleaved products. We used [GenBank: NC_001802] as a reference to know where to cut GAG and POL. In addition to evaluating our

virus-host interaction prediction method for GAG and POL cleavage products, we used GAG and POL in our analysis as well because many interactions in the NCBI database of validated HIV-human interactions are between human proteins and GAG or POL, rather than their cleavage products [62].

All protein sequences in the resulting 18 alignments, one for each HIV protein, were annotated with 133 peptide motifs (ELMs) using the ELM Resource, accessed December 2008 [131], with default settings except selecting human for the species field. Any protein lacking an ELM was removed from the study, leaving at least 70 sequences in each multiple alignment (see Supplemental table A.1). We considered an ELM to be conserved on an HIV protein if it was present on more than 70% of the protein's multiple sequence alignment. This cutoff was chosen for its stability. An increase of 5% additional conversation did not alter the number of conserved ELMs (data not shown). A total of 99 ELMs were found on at least one virus protein sequence. The conservation threshold removed 43 of these, leaving 56 total.

To assess the significance of an ELM being annotated on 70% of the protein sequences gathered for an HIV protein, we devised a control based on randomly constructed HIV protein sequences. We chose to focus on the 22 ELMs found to be conserved on the HIV Nef protein because Nef was better studied than some of the other HIV proteins [32], and had more protein sequences in the LANL HIV Sequence Database. Using our total set of 807 Nef subtype B and C protein sequences from the LANL database, we estimated the probability of amino acid occurrence as well as amino acid transition probabilities, i.e. the probability of seeing amino acid β follow amino acid α in Nef protein sequences. We constructed one random Nef protein sequence for every real Nef protein sequence by first sampling an initial amino acid based on the single amino acid probabilities, and then using the amino acid transition probabilities to sample subsequent amino acids and build the rest of the random protein sequence until it was as long as the real one.

We made one hundred sets of random Nef protein sequences, each containing

807 random Nef sequences, and matched all proteins in each set against the 133 peptide motifs from the ELM Resource. For each random protein set, we calculated the conservation of all ELMs across random Nef protein sequences in the set, and compared this conservation to the ELM conservation observed for real Nef protein sequences from the LANL database. To obtain a p-value for the conservation of an ELM on real Nef protein sequences, we recorded the number of random sets where the conservation of the ELM was equal to or greater than the conservation observed for real Nef protein sequences. We found that all of Nef's 22 conserved ELMs except the LIG_PDZ_3 motif were significantly conserved compared to random protein sequences (p-value < 0.05), i.e. all conserved ELMs but one had higher conservation on real Nef protein sequences than on random protein sequences in more than 95 of the random sequence sets. The verification that ELM conservation on Nef protein sequences was not occurring by chance made it more likely that conserved ELMs on HIV proteins were guiding interactions with human proteins.

2.2.2 Human protein peptide motif and domain annotation

Once we determined conserved peptide motifs (ELMs) for HIV proteins, we found domains (CDs) associated with these ELMs, and used them to annotate human proteins. We used the ELM Resource to find lists CDs or proteins known to interact with ELMs. For each ELM conserved on a virus protein, we found the appropriate CDs and mapped them to PROSITE domains [78]. When the ELM Resource listed a set of interacting proteins instead of CDs, we assumed that all proteins had a common unknown CD, and annotated them with that. We constructed a list of CDs and interacting proteins for each HIV conserved ELM (see Supplemental table A.2).

We annotated PROSITE domains and ELMs on the 9446 human protein sequences in the Human Protein Reference Database (HPRD) protein interaction network [129], and mapped these sequences to Entrez Gene IDs. PROSITE domains

were annotated with the PROSITE scan tool (release 20.31) using the default parameters [39]. We also annotated ELMs on human proteins using the ELM Resource, accessed August 2008, selecting the same settings used for the HIV sequences. ELMs have a tenancy to fall in regions of proteins that lack domains, and the ELM Resource uses this observation to rule out false positive ELM pattern hits on proteins [51, 119, 131]. To keep false positive hits on human proteins to a minimum, any protein lacking a PROSITE domain was removed from the study to ensure that the ELM scanner would be able to rule out some protein domain regions. After further limiting human proteins to those that interacted with one other protein in the human HPRD protein network (see next section), we were left 5954 proteins in the study.

2.2.3 Predicting interactions between HIV and human proteins

The prediction of HHP, the set of human proteins that might interact with HIV proteins, was based on interactions mediated by peptide motifs (ELMs) on virus proteins and domains (CDs) on human proteins. We built HHP from the union of two sets of human proteins, H1 and H2 (Figure 2.1). H1 was the set of human proteins predicted to directly interact with one or more HIV proteins via a human CD and a virus ELM. H2 was the set of human proteins whose interactions with proteins in H1 were potentially disrupted by competition with an HIV protein. An H1 protein has a CD that it might use to interact with an ELM present on both H2 and HIV proteins. For example, in the competition between an HIV and H2 protein for phosphorylation by an H1 kinase, the H1 protein has a kinase CD and the competing proteins have ELMs for phosphorylation sites.

The virus-host interaction prediction algorithm was straightforward. For each virus protein, we looked at all interactions documented in HPRD that could be explained by an interaction between a virus protein's conserved ELM and a CD known to interact with that ELM, and added the protein with the CD to H1 and

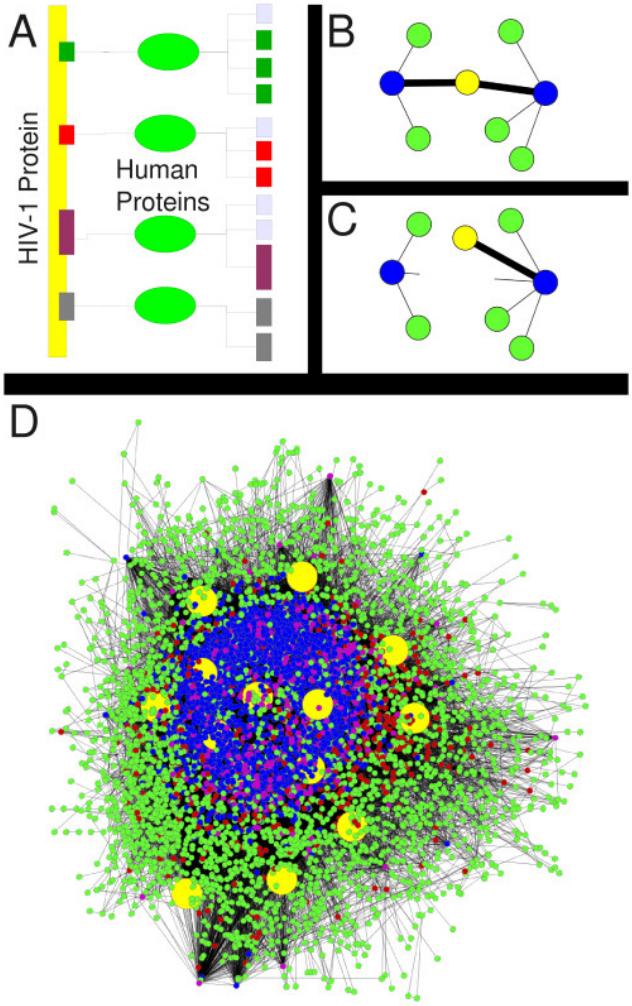


Figure 2.1: (A) The scheme for predicting HIV-human interactions. Rectangular blocks represent peptide motifs and ellipses represent the domains that interact with them. (B) An HIV protein (yellow) alters the human protein interaction network by creating a new path between human proteins (blue) and (C) breaking a path between two human proteins by competing for binding [36, 154, 168]. (D) Here we show predicted and experimentally validated HIV-human interactions with the human interaction network (HPRD). Nodes are proteins and edges represent an interaction. Yellow nodes represent HIV proteins. Purple nodes represent the overlap between predicted (HHP) and validated (HHE) interactions. Blue and red nodes represent proteins specific to HHP and HHE, respectively, while green nodes are not involved in infection.

the protein with the ELM to H2. To ensure that an interaction between H1 and H2 proteins did not involve the same human protein, we removed all such self edges from the network. Human proteins are involved in multiple protein interactions, so H1 and H2 were not mutually exclusive. H1 contained 600 proteins, H2 contained 2151, and their intersection had 403 proteins. The total set of human proteins predicted to interact with an HIV protein was the union of the HIV protein's H1 and H2 sets, and contained all host proteins that were predicted to either bind to, or compete with, the HIV protein. Across all HIV proteins, we predicted 2348 human proteins were involved in 23330 HIV-human interactions.

2.2.4 Validation using the NCBI HIV-Human Protein Interaction Database

We used the NCBI HIV-Human Protein Interaction Database (accessed August 2008), which has 3950 interactions between 19 HIV proteins and 1439 human proteins, to evaluate our HIV-human interaction predictions. HIV proteins ENV, GAG, and POL are cleaved into smaller functional proteins. The NCBI database maintains different sets of HIV-human interactions for cleavage products and the polyproteins from which they were made. For instance, HIV POL cleavage products IN, PR, and RT have some HIV-human interactions that are not attributed to POL. This distinction did not work for our evaluation purposes because under our interaction prediction method, cleavage products had a subset of the interactions attributed to their uncleaved progenitors because of the sequence overlap. For this reason, we took all ENV, GAG, and POL cleavage product virus-host interactions and assigned them to the polyprotein from which they came. When evaluating our HIV-human interaction predictions for each HIV protein, we still looked at interactions for GAG and POL cleavage products individually. However, we did not assess our predictions for human interactions with ENV cleavage products GP41 and GP120 separately from our ENV assessment because there were so few.

We restricted the human proteins interacting with HIV proteins to those that we could predict with our motif/domain prediction method, i.e. we only looked at human proteins in the HPRD human network that had domains and interacted with one human protein other than itself. These restrictions left a set of 5954 host proteins that we could predict with our algorithm. The NCBI HIV-human interactions are spread over 68 interaction types, such as 'interacts with', 'phosphorylates', and 'upregulates'. We considered all interaction types, both direct and indirect. For each HIV protein, we removed an interaction type if it described less than six interactions. This resulted in a set of 1,687 verified interactions between 15 HIV proteins and 887 human proteins. We refer to this set as HHE, and used to investigate the usefulness of our predictions. When we considered our predicted interactions for each HIV protein individually, we only looked at HIV proteins with more than ten interactions with human proteins in our restricted HPRD human network, leaving twelve HIV proteins (ENV, GAG, IN, MA, NEF, POL, PR, REV, RT, TAT, VIF, and VPR) to consider. We constructed a subset of HHE, DHHE, which had interaction types deemed to be direct by Tastan et al. [160]. DHHE was used to evaluate the portion of our predicted proteins that contained domains, as these proteins were more likely to have direct interactions with HIV proteins.

Computations used for the comparison of predicted and validated HIV targeted human proteins

To compare our predicted set of interactions with the experimentally verified dataset from NCBI, we focused on the overlap between the two sets, Gene Ontology (GO) [7] molecular function enrichment, GO biological process similarity, and KEGG pathway [85] enrichment. P-values for the overlap between sets of predicted and verified HIV targeted proteins and their various subsets were calculated using the hypergeometric test using a background set of 5954 possibly predicted human proteins. P-values for GO and KEGG term enrichment for a given protein set compared to the background

set of 5954 possibly predicted human proteins were found using Bonferroni corrected p-values from the Database for Annotation, Visualization, and Integrated Discovery (DAVID) tool [43].

To statistically compare the GO biological process similarity between predicted and validated virus targeted proteins, we devised a permutation test based on GO similarity between proteins, as calculated by the GS2 tool [142]. GO is a hierarchy of labels, with general biological processes at the first level, and more specific ones on higher levels. The GS2 tool takes two proteins, finds where their GO labels are in the GO hierarchy, and calculates a distance between all GO labels based on how far away they are in the GO hierarchy. The GO similarity between two proteins is the average GS2 distance between the GO labels that annotate the proteins. When we compared predicted and validated virus targeted proteins, we limited the comparison to only proteins with GO biological process labels in the fifth level of the GO hierarchy. We chose this level because the biological process labels here are specific enough for a meaningful comparison, yet general enough to annotate large numbers of proteins in our predicted and validated protein sets.

We performed the test for significant GO similarity between predicted and validated HIV targeted proteins for HIV proteins with at least ten targeted human proteins with GO labels in the fifth level of the GO hierarchy. To find the GO similarity between predicted and validated protein sets, we first found the GO similarity between all protein pairs taken from the two sets, and then averaged the GO similarity for all cross set comparisons. For the permutation test to assess the significance of the GO similarity between predicted and validated virus targeted protein sets, we constructed one hundred random sets of predicted human targets for each HIV protein by sampling from a background set of 4501 human proteins that were both considered in our study of HPRD human network proteins, and had GO biological process labels in the fifth level of the GO hierarchy. For each HIV protein's GO similarity between predicted and verified human targets, we arrived at a p-value

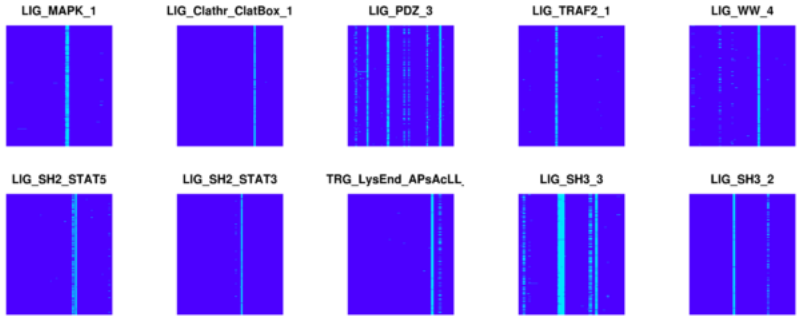


Figure 2.2: This figure shows host peptide motif conservation on NEF. Peptide motifs (ELMs) were spatially conserved on alignments of HIV proteins of subtypes B and C. Each box shows the annotations for one conserved ELM (present on more than 70% of protein instances) on the multiple alignment of NEF protein sequences taken from different patients. An ELM can be spatially conserved in multiple positions on the alignment, demonstrated by multiple sets of thick vertical lines in an ELM's box.

describing the significance of the similarity by counting the number of random predicted sets that had an equal or greater GO similarity with the validated HIV-human interactors.

2.3 Results

2.3.1 Human peptide motifs were conserved on HIV proteins

Figure 2.2 shows a subset of the conserved peptide motifs (ELMs) annotated on HIV NEF's multiple protein sequence alignment. It is clear from the figure that conserved ELMs occur in roughly the same position on each aligned protein. Our computations showed that this was true for all conserved ELMs on all HIV proteins. The HIV reverse transcription process is susceptible to errors, with HIV RT making roughly 0.2 errors per genome in each replication cycle [134], which leads to an evolutionary rate one million times that of host genomes [108]. Noting that HIV is

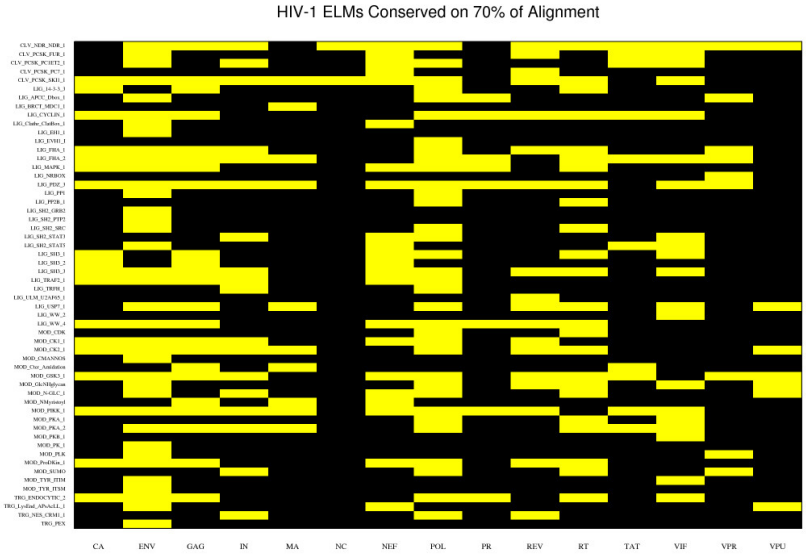


Figure 2.3: Here we show conserved host peptide motifs (ELMs) on HIV proteins. Of 133 ELMs scanned, only 56 were conserved (present on more than 70% of an HIV protein's sequence alignment). Yellow boxes indicate conservation of an ELM above 70% for a virus protein. All HIV proteins shown had at least one conserved ELM.

a virus with high mutation rate, it is likely that conserved ELMs are essential for viral replication within the host cell [82]. ELM annotation in eukaryotic proteomes is not yet complete. Multiple computational strategies have been employed for the discovery of additional ELMs involved in protein interactions and post-translational modifications [45, 120]. It is possible that HIV proteins have additional conserved ELMs that have not yet been identified.

Conserved peptide motifs (ELMs) are shown for each HIV protein in Figure 2.3. Overall, 56 of the 133 ELMs in the ELM Resource were conserved on some HIV protein. Some of the conserved ELMs, like the SH3 ligand sites on NEF, have been experimentally verified as binding sites for human proteins [32]. We found that some conserved ELMs occur frequently on human proteins. A PDZ domain binding site, ELM LIG_PDZ_3, was seen on 90% of human proteins. Other ELMs, like LIG_EH1_1, a binding site for the WD40 domain, appeared on only a few human proteins (see

Supplemental table A.1).

2.3.2 Significant overlap of predicted and validated HIV-human interactions verifies our method

The NCBI set of curated HIV-human interactions contains 887 host proteins known to interact with one or more HIV proteins. The dataset captures both direct and indirect, regulatory HIV-human protein interactions [62], and was appropriate for the task of assessing our predicted interactions because it allowed us to judge our algorithm's ability to capture both direct and indirect interactions. The HPRD human protein interaction network containing the 5954 human proteins in this study is shown in Figure 2.1D with yellow HIV proteins connected to their predicted interaction partners (blue) and their verified interaction partners (red). Proteins in both sets are purple, while all other proteins are green. As seen in the figure, our predicted set of virus targeted human proteins, with over two thousand proteins, was larger than the verified NCBI list, with only 877 proteins.

For a more quantitative evaluation of our predictions, we compared predicted and verified HIV targeted proteins for all HIV proteins individually. The significance of the overlap between predicted and verified interactions is shown in Table 2.1. Of the twelve virus-host predicted interaction sets we evaluated, only two, from HIV proteins IN and VIF, did not have significant overlap with the NCBI interactions. While the overlap between predicted and validated interactions was significant (p-value < 0.05), and the recall of known virus-host interactions for each HIV protein was at least 20%, there were many predicted interactions that were not validated by the NCBI database.

Our predicted interactions consisted of human proteins with domains that interact with conserved HIV peptide motifs (H1 proteins), and human proteins that that interacted with H1 proteins and were annotated with HIV peptide motifs (H2 proteins). Proteins in H2 dominated the overlap between predicted and validated

HIV protein	HHP	HHE	Overlap	P-value
ENV	2166	409	194	8.09e-07
GAG	2035	103	46	9.94e-03
IN	1759	46	12	0.631
MA	1129	47	19	1.60e-04
NEF	1828	155	83	6.42e-10
POL	2093	122	57	2.96e-03
PR	1169	58	20	2.30e-03
REV	1702	40	16	4.11e-02
RT	1986	23	20	1.07e-08
TAT	1106	509	183	5.54e-23
VIF	1832	35	7	0.888
VPR	919	119	39	5.29e-07

Table 2.1: For each HIV protein, we evaluated the significance of the overlap between human proteins in HIV-human predicted and validated interactions. The HHP column gives the number of human proteins that were predicted to interact with an HIV protein, while HHE shows the number of verified virus targeted proteins in the NCBI database. The Overlap column counts the number of proteins in both sets. P-values for the overlap between predicted and validated HIV targeted proteins were calculated using a hypergeometric test (see Methods). We limited our results to HIV proteins with at least ten verified interactions with human proteins.

virus targeted proteins. Roughly two thirds of the proteins in H1 were also found in H2. For this reason, we sought to evaluate H1 separately from our total prediction set. For each HIV protein, we investigated the usefulness of H1 by comparing it with DHHE, the validated NCBI direct virus-host interactions. We limited our comparisons to the twelve HIV proteins with at least ten direct interactions with human proteins, and found that eight of these HIV proteins (ENV, GAG, MA, NEF, POL, REV, RT, and TAT) had significant overlap between H1 and DHHE (see Supplemental table A.3). Figure 2.7 shows the overlap p-values and sizes of DHHE and H1 for HIV proteins ENV, NEF, and TAT. Our H1 predictions for HIV proteins IN and VIF failed to have significant overlap with validated direct interactions, just as our full set of predicted interactions for these HIV proteins did not have significant overlap with all verified virus-host interactions (Table 2.1). While for the HIV RT protein, H1 predictions performed better than the full prediction set (HHP), this

VP	H1	DHHE	Match	Precision	Recall	Pvalue
ENV	527	177	25	4.74	14.12	6.57E-003
NEF	399	55	12	3.01	21.82	4.92E-005
TAT	188	213	24	12.77	11.27	8.93E-009

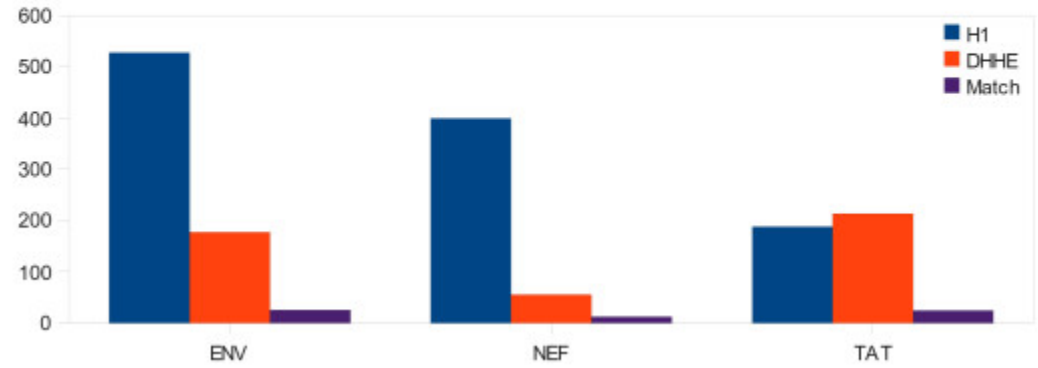


Figure 2.4: Here we compare predicted and validated virus-host direct interactions for HIV proteins ENV, NEF, and TAT. The Match column holds the overlap between predicted and verified virus targeted host protein sets. The figure compares host proteins from direct interaction predictions (H1) with host proteins from experimentally verified direct interactions (DHHE). The p-values indicated a significant overlap for all protein sets. P-values were calculated as described in Methods.

was not the case for the other HIV proteins. For both H1 and HHP prediction sets, our precision was not high, indicating that most proteins predicted to interact with HIV were not verified by the NCBI database. For this reason, we turned to the Gene Ontology and the KEGG pathway databases for an understanding of how our unverified predictions captured the host cell biological pathways and functions targeted by HIV proteins.

2.3.3 Predicted and validated HIV-human interactions share similar Gene Ontology labels

As a biological validation of our results, we compared our predicted and validated HIV-human interactions using Gene Ontology (GO) labels of molecular function and biological process. While our predictions missed some of the exact interactions with

Gene Ontology Label	ENV HHP/HHE	NEF HHP/HHE	TAT HHP/HHE
adenyl ribonucleotide binding	6.82e-12/0.003	NA /NA	NA/NA
inositol or PI kinase activity	7.22e-07/0.0009	NA/NA	1.033e-06/0.0016
lipid kinase activity	7.22e-07/0.0009	NA/NA	2.06e-08/0.0016
MAP kinase activity	0.00029/2.33e-07	NA/NA	NA/NA
phosphoinositide 3-kinase activity	0.0002/0.0001	NA/NA	1.54e-05/0.000167
protein kinase activity	5.60e-34/0.00085	2.04e-29/0.00044	1.01e-36/0.0035
protein kinase binding	NA/NA	NA/NA	2.16e-18/3.46e-06

Table 2.2: Here we show Gene Ontology molecular function level 5 labels statistically enriched (p-value < 0.01) on human proteins from our predicted virus-host interactions (HHP) for HIV ENV, NEF, and TAT. Enrichment for host proteins involved in NCBI's verified virus-host interactions (HHE) is also indicated.

HIV proteins, they might have been close to them in terms of molecular function and biological pathways. GO labels are organized hierarchically, with general labels towards the first levels of the hierarchy, and specific labels at the higher levels. For our investigations, we focused on specific terms on the fifth level of the GO hierarchy. We examined GO labels for predicted and validated virus-host interactions for HIV proteins ENV, NEF, and TAT because they had the most verified experimental interactions with human proteins. First, we found the GO molecular function level 5 labels that were enriched in our predicted virus targeted human proteins, and then calculated the enrichment for the host proteins in the validated NCBI HIV-human interactions. Table 2.2 shows that GO molecular functions enriched on predicted virus targets were also enriched on validated HIV targets (p-value < 0.01).

For a more quantitative assessment of our predicted HIV-host interactions, we measured the similarity between predicted HIV targeted pathways and known HIV targeted pathways, using GO biological process level 5 labels as a proxy for pathway annotations. For each HIV protein, we computed the average GO similarity for all predicted/validated protein pairs taken from host proteins in predicted and validated virus-host interactions, and found the significance of the GO similarities using a permutation test (see Methods). Table 2.3 shows that predictions for most HIV proteins had significant GO similarity with verified virus targeted proteins.

HIV protein	GO HHP	GO HHE	P-value
ENV	1640	347	0.00
GAG	1545	88	0.00
IN	1347	44	0.99
MA	872	45	0.00
NEF	1376	126	0.00
POL	1595	116	0.00
PR	882	54	0.00
REV	1273	35	0.00
RT	1517	23	0.00
TAT	857	445	0.00
VIF	1397	32	0.98
VPR	694	93	0.00

Table 2.3: For each HIV protein, we evaluated the performance of our HIV-human interaction predictions (HHP) using Gene Ontology (GO) biological process labels to compare our predicted interactions to experimentally validated interactions. We computed a GO similarity score between predicted and validated protein sets by averaging over all GO similarity scores that resulted from pairwise combinations of proteins taken from the predicted and validated protein sets (see Methods). For each HIV protein, we constructed random HHP sets to use in a permutation test to evaluate the significance of the observed GO similarity between HHP and HHE. We report the p-values for these trials.

Just as predictions of virus-host interactions for HIV proteins IN and VIF did not show a significant overlap with verified interactions, the GO similarities between predicted and verified virus-host interactions for these HIV proteins were not found to be statistically significant. For the other HIV proteins, it is likely that some of our unverified predictions from Table 2.1 and Figure 2.4 are correct, or have some importance for HIV infection, because they act in the same biological processes that HIV targets.

2.3.4 Predicted and validated HIV-human interactions occupy the same KEGG pathways

Since our predicted virus-host interactions performed well at recovering known HIV targeted biological processes, we moved to an evaluation of our predictions that involved more defined biological pathways from the KEGG pathway database. The

KEGG Pathway	ENV HHP/HHE	NEF HHP/HHE	TAT HHP/HHE
AML	1.97E-05/2.68E-05	2.07E-05/8.55E-05	2.66E-08/3.08E-04
Adherens junction	2.15E-09/NA	8.21E-11/NA	6.74E-06/NA
Apoptosis	4.58E-04/1.43E-13	5.18E-04/6.54E-04	3.91E-03/3.88E-13
B cell receptor signaling	4.86E-06/8.73E-10	2.88E-04/2.90E-04	1.25E-09/5.57E-06
Cell cycle	NA/NA	NA/NA	2.88E-04/1.90E-01
CML	7.21E-05/2.23E-05	1.17E-06/2.46E-05	1.01E-09/7.53E-07
Colorectal cancer	1.07E-05/3.18E-04	4.68E-08/3.31E-03	4.50E-04/1.42E-03
Endometrial cancer	6.03E-04/1.62E-02	1.66E-04/2.10E-03	5.03E-07/3.83E-04
H. pylori infection	2.29E-04/3.84E-06	2.07E-06/2.83E-05	NA/NA
ErbB signaling	2.27E-10/5.70E-04	1.70E-12/9.22E-04	1.53E-12/1.66E-05
Fc epsilon RI signaling	8.43E-04/6.23E-21	2.14E-05/2.20E-07	9.62E-05/1.52E-04
Focal adhesion	2.82E-06/2.30E-03	2.31E-07/6.90E-02	5.28E-08/3.36E-09
Gap junction	1.90E-04/1.26E-04	NA/NA	1.12E-04/7.18E-10
Glioma	1.50E-04/1.24E-05	4.99E-06/6.02E-03	1.56E-07/6.79E-11
Insulin signaling	1.53E-07/8.84E-02	1.73E-04/3.50E-01	2.91E-07/7.35E-02
Jak-STAT signaling	4.08E-08/2.15E-04	4.09E-09/1.28E-01	2.32E-17/4.91E-03
Leukocyte migration	1.94E-07/2.21E-08	1.17E-08/6.36E-01	3.28E-05/1.45E-01
Long-term potentiation	6.79E-05/2.20E-02	NA/NA	9.04E-03/2.38E-10
MAPK signaling	5.19E-08/6.32E-04	1.58E-09/3.27E-03	1.18E-03/5.15E-01
NK cell cytotoxicity	NA/NA	NA/NA	9.50E-06/5.31E-15
Non-small cell lung cancer	4.28E-05/1.25E-04	1.26E-05/1.67E-03	7.55E-06/1.45E-06
Pancreatic cancer	1.26E-04/5.54E-07	1.10E-05/2.50E-06	1.03E-05/8.15E-08
E. coli infection	3.77E-03/1.00E+00	2.94E-03/NA	9.74E-03/3.25E-01
PtdIns signaling	1.36E-03/1.72E-04	2.37E-03/NA	2.19E-05/9.74E-06
Prostate cancer	1.90E-04/1.26E-04	6.26E-06/6.56E-05	5.46E-09/1.11E-07
Regulation of cytoskeleton	4.30E-03/6.02E-01	1.73E-03/8.79E-01	2.66E-03/7.65E-01
Small cell lung cancer	1.94E-03/3.71E-10	8.42E-05/4.25E-02	1.12E-04/4.09E-14
T cell receptor signaling	NA/NA	NA/NA	1.56E-06/1.35E-11
Tight junction	1.24E-03/1.00E+00	5.29E-04/NA	NA/NA
Toll-like receptor signaling	5.16E-03/2.04E-14	5.37E-05/2.04E-14	NA/NA
Type II diabetes mellitus	NA/NA	NA/NA	3.47E-03/5.95E-01
VEGF signaling	3.23E-03/4.89E-15	6.79E-03/8.82E-03	1.88E-05/4.07E-12

Table 2.4: Here we show KEGG pathways enriched (p-value < 0.01, see Methods) with human proteins from our predicted virus-host interactions (HHP) for HIV ENV, NEF, and TAT. Enrichment for host proteins involved in NCBI's verified virus-host interactions (HHE) is also indicated.

KEGG pathways statistically enriched for HIV ENV, NEF, and TAT interacting proteins (experimental as well as computational) included immune system pathways such as T cell and B cell receptor signaling pathways, apoptosis, focal adhesion, and toll-like receptor signaling pathways (Table 2.4). Gene expression data before and after HIV infection of macrophages also showed apoptosis and MAPK signaling pathways as statistically enriched [20], as predicted here. Microarray results did not show cell cycle and toll-like receptor pathways as highly activated in HIV activated macrophages, although the toll-like receptor pathway was highly enriched with known HIV targeted proteins (Table 2.4). Also statistically enriched were disease pathways such as the colorectal cancer, leukemia, and lung cancer pathways that have been shown to have high incidence of occurrence in HIV infected individuals [127]. Other disease pathways predicted by our analysis included those previously associated with HIV infection: H. pylori infection [124], E. coli infection [133], and type II diabetes [125]. These observations indicated the promise of our method in predicting activated disease pathways based on viral sequence. Post-translational modification appeared to be an important element of HIV cellular network hijacking. As shown in Table 2.2, protein kinase activity and protein kinase binding were significantly enriched both in predicted and verified HIV targeted proteins, suggesting the importance of altered phosphorylation events in the reorientation of the host cell interaction network towards virus survival and replication [20]. The HIV activated GO categories listed in Table 2.2 are associated with signal transduction processes in the KEGG pathways presented in Table 2.4.

The positions of predicted and matched HIV targeted proteins along KEGG pathways allowed us to assess the overlap between computational and experimental prediction based on cell-compartment identity. Figure 2.5 shows the overlap (purple) between predicted (blue) and experimentally determined (red) host proteins targeted by HIV TAT along the natural killer cell mediated cytotoxicity pathway.

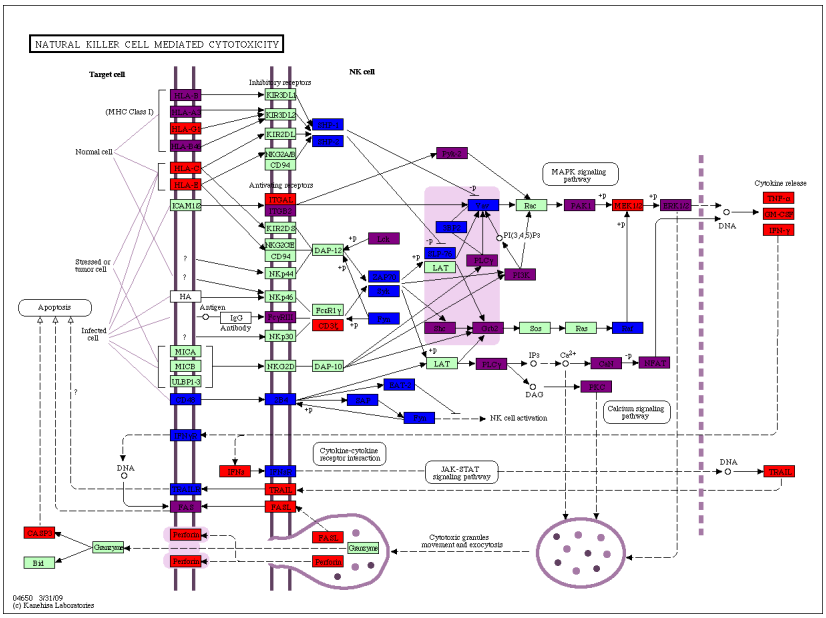


Figure 2.5: HIV TAT natural killer cell mediated cytotoxicity. The KEGG natural killer cell mediated cytotoxicity pathway is colored for predicted TAT-human interactions (blue) and validated TAT-human interactions (red), and their overlap (purple). Green boxes have proteins not involved in infection, while white boxes do not have human proteins.

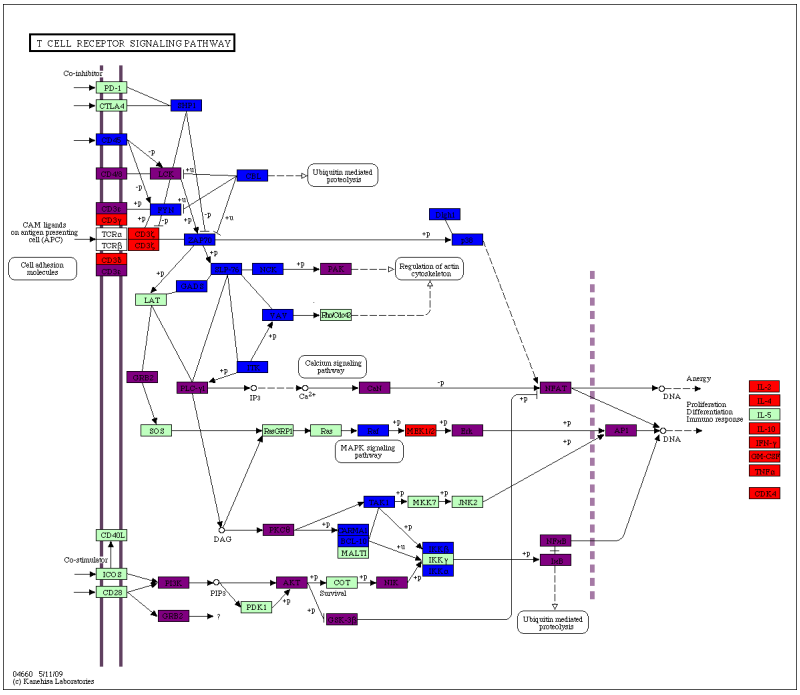


Figure 2.6: HIV TAT T cell receptor signaling pathway. The KEGG T cell receptor signaling pathway is colored for predicted TAT-human interactions (blue) and validated TAT-human interactions (red), and their overlap (purple). Green boxes have proteins not involved in infection, while white boxes do not have human proteins.

Our predictions were on target on the cell membrane for HLA-B, HLA-A3, HLA-B45, and FAS, but we missed Perforin, HLA-C, HLA-E, and HLA-G1. The figure also shows a good match for DNA transcription factors targeted by HIV. The green boxes in the figure correspond to host proteins with apparently no direct interaction with TAT. The T cell receptor signaling pathway in Figure 2.6 had a high degree of overlap (purple) along the cell membrane and on transcription factors between TAT targeted host proteins (red) and our corresponding predictions (blue). The abundance of predicted host proteins in the pathway with no matching experimental result suggested new virus-host interaction studies for HIV as well as a justification for further refinement of our computational method, perhaps by incorporating protein structures [46].

Human proteins in our predicted HIV-human interactions performed well at capturing virus-targeted KEGG pathways. One reason that this pathway evaluation showed our predictions to be more promising than a simple protein set comparison could be that proteins in KEGG pathways are better studied than proteins in the HPRD human interaction network [85]. We attempted to obtain a set of predicted interactions with fewer false positives by limiting human proteins in our predicted interactions to those that were in KEGG pathways. We then compared the overlap between KEGG restricted predicted virus-host interactions and validated virus-host interactions for HIV proteins ENV, NEF, and TAT (Figure 2.7). We found that the intersection between predicted and verified virus-host interactions for human proteins in the HPRD human interaction network became more significant as we limited predictions to proteins in KEGG pathways (Figure 2.7 and Supplemental table A.4). Restricting predictions to all KEGG pathways produced a set of virus-host interactions with less false positives, so we decided to restrict our predictions further by finding KEGG pathways that were enriched with human proteins from predicted HIV-human interactions, and keeping only predictions in these pathways.

Viruses often target specific host pathways, like the type I interferon response pathway, and interact with multiple host proteins in these pathways [116, 181]. We hypothesized that estimating HIV targeted pathways from our predicted virus-host interactions, and then restricting our predictions to only human proteins in these pathways would yield better results than limiting our predictions to all KEGG pathways. We were further motivated to find KEGG pathways enriched with our predictions because studying proteins in specific pathways is valuable because predicted interactions with certain pathways can be used to guide targeted virus-host interaction experiments [96]. We limited our predicted interactions by focusing on human proteins in KEGG pathways that were found to be enriched with our predictions (p-value < 0.01, see Methods). Bar graphs in Figure 2.7 demonstrate the intersection of predictions when restricted to the KEGG pathways in which they are enriched

with verified interactions for HIV ENV, NEF, and TAT. Compared to limiting our predictions to proteins in all KEGG pathways, restricting predictions for HIV ENV and NEF to proteins in KEGG pathways that were statistically enriched with human proteins in our virus-host predicted interactions for these HIV proteins improved the overlap between predicted and verified virus-host interactions.

2.3.5 Virus level comparison of predicted and verified HIV-human interactions

We have evaluated our predicted interactions for individual HIV proteins, but another way to view our predictions is at the virus level. At the virus level, we can check to see if human proteins that were predicted to interact with one HIV protein have any validated interactions with other HIV proteins. This test was motivated by the observation that virus proteins often interact with the same host proteins [130]. Figure 2.8 shows a combined view of predicted and validated virus-host interactions, made by aggregating interactions for all virus proteins. When we restricted our predictions to KEGG proteins, we had 1047 host proteins, and 345 of these had already been shown to be interacting with at least one HIV protein. The match between computational prediction and experimental data in this case led to a p-value of 1.97e-62. The improvement seen in recall and precision for virus level predictions compared to HIV protein level predictions indicated that our predictions captured many of the host interactions shared between HIV proteins, and that interaction predictions made for one virus protein could be used to capture interactions with other HIV proteins.

Projection	VP	HHP	HHE	Match	Precision	Recall	Pvalue
All Pathways	ENVa	1013	409	170	16.78	41.56	5.48E-035
HHP Enriched Pathways	ENVe	584	409	127	21.75	31.05	7.70E-037
All Pathways	NEFa	866	155	69	7.97	44.52	1.44E-020
HHP Enriched Pathways	NEFe	519	155	54	10.4	34.84	4.26E-021
All Pathways	TATa	621	509	150	24.15	29.47	2.96E-037
HHP Enriched Pathways	TATe	410	509	112	27.32	22	1.57E-032

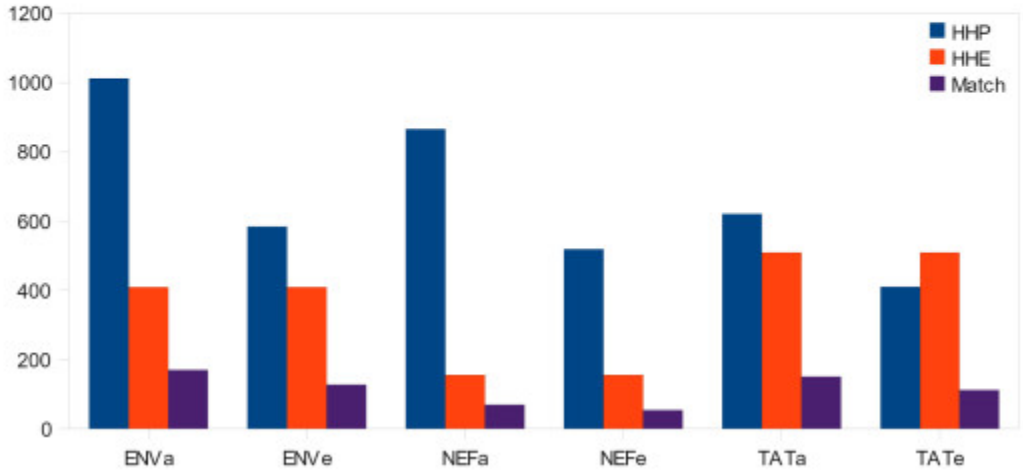
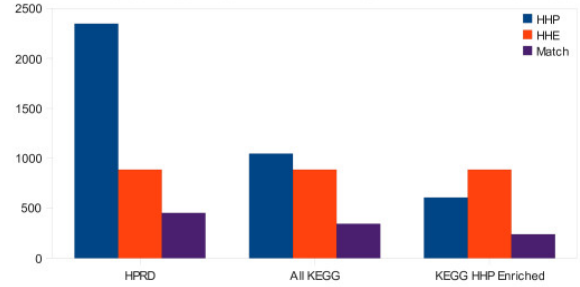


Figure 2.7: Here we compare predicted and validated virus-host interactions for host proteins in KEGG pathways. The Match column holds the overlap between predicted and verified virus targeted host protein sets. The figure compares host proteins from all predicted (HHP) and verified (HHE) interactions for the three HIV proteins. Predicted host proteins were restricted to either genes in all KEGG pathways (ENVa, NEFa, TATa), or KEGG pathways enriched (p-value < 0.01, see Methods) with our predictions (ENVe, NEFe, TATe). The intersection between predicted and verified interactions was significant for both restrictions, but slightly more significant for enriched pathways for HIV proteins ENV and NEF. P-values were calculated as described in Methods.

Projection	HHP	HHE	Match	Precision	Recall	Pvalue
HPRD	2348	887	454	19.34	51.18	5.72E-015
All KEGG	1047	887	345	32.95	38.9	1.97E-062
KEGG HHP Enriched	607	887	241	39.7	27.17	3.11E-058



	HHE	HHP	Konig	Brass	Zhou
HHE	887	5.72E-015	0.01	2.20E-004	0.05
HHP	454	2348	0.15	0.03	0.52
Konig	29	56	129	7.35E-004	0.36
Brass	26	44	7	91	4.46E-005
Zhou	27	54	3	9	139

Figure 2.8: Here we evaluate our predictions on the virus level, rather than the virus protein level. For both predicted and validated virus-host interactions, we combined the host targets of individual HIV proteins to produce virus level protein sets. The overlap (Match) of our predictions (HHP) with verified HIV targeted proteins (HHE) was compared when restricting them to proteins in HPRD, KEGG, and KEGG pathways enriched in HHP (p-value < 0.01, see Methods). The lower table compares HHE, HHP, and predictions from three siRNA screens. The darkened diagonal holds the sizes of all sets. The overlap between sets is below the diagonal, while p-values for these overlaps are above (see Methods).

ELM Type	VP	HHP	HHE	Match	Precision	Recall	Pvalue
Conserved ELMs	ENVc	2166	409	194	8.96	47.43	8.09E-007
Frac .25 Win 20	ENV.25	1388	409	138	9.94	33.74	2.43E-007
Frac .50 Win 20	ENV.5	1989	409	175	8.8	42.79	1.65E-005
Conserved ELMs	NEFc	1828	155	83	4.54	53.55	6.42E-010
Frac .25 Win 20	NEF.25	943	155	52	5.51	33.55	9.40E-009
Frac .50 Win 20	NEF.5	1417	155	65	4.59	41.94	1.34E-007
Conserved ELMs	TATc	1106	509	183	16.55	35.95	5.54E-023
Frac .25 Win 20	TAT.25	181	509	50	27.62	9.82	3.52E-015
Frac .50 Win 20	TAT.5	643	509	126	19.6	24.75	1.14E-021

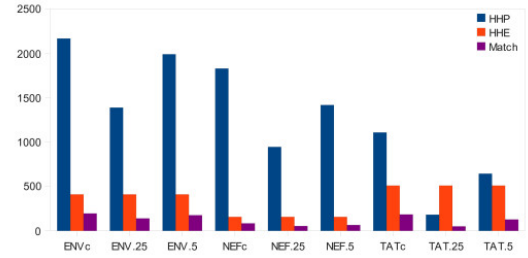


Figure 2.9: Here we tested the hypothesis that using infrequent host peptide motifs for HIV-human interaction prediction, rather than all host peptide motifs conserved on HIV proteins, improved the performance of our predictions. We found infrequent host peptide motifs by limiting conserved HIV peptide motifs to those that occurred on less than some fraction (Frac) of human proteins, or were seen on an HIV protein with another peptide motif within a twenty residue window, resulting in a motif module. For HIV proteins ENV, NEF, and TAT, we compared the performance of predictions using two human fraction cutoffs, 0.25 and 0.5, to predictions made with unrestricted conserved HIV peptide motifs (Conserved ELMs). We found significant overlap (Match) between predicted and validated virus-host interactions in all cases, but using the fraction cutoffs only helped for the HIV ENV protein. P-values were calculated as described in Methods.

2.3.6 Infrequent host peptide motifs did not improve prediction performance

We next asked if our predictions could be improved by limiting spurious peptide motif pattern hits on host proteins. Some peptide motifs, like the PDZ ligand LIG_PDZ_3, have patterns that match 90% of host proteins (Supplemental table A.1). We hypothesized that peptide motifs, or ELMs, that occurred infrequently in the host proteome would have a higher chance of being functional than frequently occurring ELMs. Capturing more functional ELMs should result in better HIV-human interaction predictions. We restricted our conserved virus ELMs to those with infrequently occurring host pattern matches in two ways. First, we imposed a frequency cutoff based on the fraction of host proteins annotated with the ELM pattern. Second, we looked for ELM modules, defined as two different ELMs occurring in a 20 residue window. ELMs often occur as modules within the same region of proteins, acting in a concerted and cooperative fashion, or as regulatory switches [65]. Since functional ELMs often occur in together, limiting ELMs to those occurring in modules decreased the false positive pattern matches for ELMs.

We identified ELM modules conserved on more than 70% of each HIV protein's multiple sequence alignment, as we did for ELMs. We found the fraction of human proteins with each ELM or ELM module, and chose two fraction cutoffs, 0.25 and 0.50, to restrict the ELMs and ELM modules on virus proteins to those that were infrequent on human sequences. Any ELM or ELM module with a human frequency above the cutoff was not used to predict interactions. Figure 2.9 shows the results for HIV proteins ENV, NEF, ad TAT, and compares the use of all conserved ELMs to using frequency (fraction) cutoffs for conserved ELMs and ELM modules. The results indicated that such restrictions on ELMs helped results for ENV, but not for NEF and TAT. For NEF and TAT, ELM restrictions yielded smaller set of predicted interactions, but the overlap between predicted and verified virus-host interactions was also reduced.

2.4 Discussion

The rapid sequencing of viral genomes with next generation sequencing technology [94] makes it possible to link clinical parameters of viral infection to peptide sequence motifs. The task of identifying host proteins targeted by a virus is worthwhile because such proteins may become drug targets to fight infection [19] and guide experiments [79, 96]. Experimental studies for determining virus targeted proteins are expensive and highly challenging [82]. Such efforts, although large-scale, have produced incomplete results for even well studied viruses like HIV [19, 92, 184]. In this study, we used a systems approach to identify host protein subsets enriched by virus targeted proteins. Our method was based on the identification of host peptide motifs on virus protein sequences. We used the a priori knowledge in the ELM Resource to identify the counter domains associated with these peptide motifs, and information from the human protein interaction network to focus on host protein interaction pairs with appropriate motif/domain links. KEGG pathways and the GO annotations were used to provide biological context and validation for our predicted virus-host interactions.

The sets of host proteins we predicted as targeted by a given HIV protein in KEGG pathways were statistically enriched with host proteins known to interact with the same HIV protein (Figure 2.7). For example, the match between our predictions and the interactions for HIV NEF in the NCBI HIV-Human Protein Interaction Database corresponded to a p-value of 4.26e-21 in KEGG pathways enriched in our predicted set. After combining our predictions for all HIV proteins, we had 607 proteins in HHP enriched KEGG pathways, and of these we matched 241 in the set of 877 experimentally verified proteins with a p-value of 3.11e-58 (Figure 2.8). Our predictions were not nearly an exact match for experimental data, but our list was highly enriched with HIV targeted host proteins. Given that reducing our total virus-host predictions to those with human proteins in KEGG pathways removes roughly half of the interactions, and has a stronger overlap with verified

virus-host interactions, experimentalists should begin verification with this set using rapid testing of binary interactions, such as yeast or mammalian two-hybrid assays [158], or protein fragment complementation assays [136].

In addition to the interaction research compiled in the NCBI HIV-Human Protein Interaction Database, recent experimental studies based on genome-wide small interfering (siRNA) screens have brought additional light to host-pathogen interactions that facilitate HIV replication [19, 92, 184]. In these screens, host genes were knocked down in HIV infected cells, and the effect on the virus is observed. Genes whose expression depletion negatively affected the virus were recorded as hits, i.e. host factors that are required for HIV replication [24]. Three siRNA studies produced smaller lists of host proteins than the list in the NCBI HIV-Human Protein Interaction Database. The lower matrix in Figure 2.8 shows the five-way comparison of HIV targeted protein lists: verified HIV targeted human proteins from NCBI, human proteins in our predicted virus-host interactions, and the three siRNA screens.

Figure 2.8 indicated the extent of discrepancy between lists, as well as the statistical significance of the overlap between them. Our predictions matched the validated NCBI virus-host interactions with the lowest p-value, and the genome-wide study lists generally matched each other better than the interaction studies. The list of 280 genes presented as host cellular factors required for HIV replication by Brass et al. had 13 genes in common with the list of 295 genes deemed necessary by Konig et al. for regulation of early stage HIV replication, and shared 10 genes with the 311 genes given in the Zhou study. When these proteins were limited to proteins in the HPRD human protein interaction network, the overlap between them led to p-values of 7.35e-4 and 4.46e-5. Although the overlap was significant, there was still a discrepancy between the results. This mismatch may be attributed to the differences in the analysis and experimental methodologies used [24]. Our predictions matched 56 of the 129 HPRD proteins presented by Konig et al. with a p-value of 0.15, 44 of the 91 HPRD proteins in the list by Brass et al. with a p-value of 0.03, and 54 of

the 139 HPRD proteins given by Zhou et al. with a p-value of 0.52. The significant overlap between our human proteins in our predicted virus-host interactions and the Brass et al. screen is promising for a focused experimental study of virus-host interactions that facilitate HIV replication. Such a study would only experimentally test predicted virus-host interactions where the human protein has been implicated in an siRNA screen. Since our predicted interactions are guided by motifs on virus proteins and domains on human ones, those that are verified experimentally already have proposed protein binding regions that could be targeted with drugs [126].

Although our study produced host protein sets statistically enriched with proteins known to be targeted by HIV, mismatches between our predictions and experimental data cannot be ignored. It is possible that virus-host interactions are guided by sequence features more complex than the peptide motif and domain interactions used in this study. The molecular vocabulary of protein interactions is simply not well understood even for proteins belonging to the same species. However, one common mode of interaction is the binding of a peptide motif on one protein to a domain on another protein [114]. A central hypothesis in the discovery of the linear binding motifs mediating protein interactions has been that proteins with a common interacting partner, such as protein kinases, share a common feature in the form of a motif [118]. Some of the peptide motifs in the ELM Resource have been shown to bind directly to sites at opposing counter domains listed in databases such as PROSITE and Pfam [117]. However, for approximately 30% of the protein interactions listed in HPRD human interaction network, interacting proteins possess none of the already annotated domains. Thus, a model based on known motif/domain interactions would not be able to capture all of the known interactions in the host, let alone those between virus and host.

Another important cause of the discrepancy between our predictions and experimental data might have been the poor annotation of known motifs and domains used in this study [44]. Recent studies of domain-motif interactions indicated that

the domains can be divided into more specific versions than those presented Pfam and PROSITE. This was found to be true for the HIV interacting PDZ domain [166], SH3 domain [149] and others [82]. Based on these and other observations, a new database of revised Pfam domains is currently under development (Robert Weatheritt, personal communication, March 8, 2010).

Emerging peptide motif discovery tools will help researchers improve the specificity of the motifs that mediate virus-host interactions, a task which is difficult because motifs are small and variable. As few as two sites in a peptide motif may be important for activity [51]. The assumption made by peptide motif discovery tools is that proteins that interact with a given protein will have over-represented peptide motifs that cause their common interaction with the given protein [159]. The Discovery of Linear Motifs (DILIMOT) server finds over-represented peptide motifs in a set of query proteins, scoring motifs by the number of query sequences with the motif, the lengths of the query sequences, and the conservation of the motif among known orthologs [119]. Like the ELM Resource, DILIMOT does not search for peptide motifs in protein domains. An alternative small linear motif (SLIM) detector, SLIMFinder, constructs over-represented motifs by combining dimers of residues to form longer patterns, and retains only those motifs occurring in a sufficient number of unrelated proteins [51]. While our prediction method could be improved with more knowledge of protein motifs and domains, the list of host proteins we have provided
(http://www.biomedcentral.com/content/supplementary/1755-8794-2-27-s5.xls) comprises a candidate set for genome-wide studies of the regulation of HIV replication and infection.

We focused on HIV infection in this study because we desired to assess the effectiveness of our computational approach by comparing our predictions with large-scale experimental data. Our results provided a rationale for applying our method to predict virus-human interactions for sequenced viruses. A future systems approach to

predicting host-pathogen interactions will at least be partially based on the sequence motifs of interacting genome/proteomes. The present study illustrated the importance of peptide motifs in the molecular cross talk between host and virus and opened the door for more extensive experimental and computational studies of virus-host interactions.

2.5 Conclusion

In this study, we described a bioinformatics model to investigate the interactions between the HIV and human proteins. Our method used multiple sequence alignments of HIV proteins, and three datasets related to the host: sequences of the host proteins, a priori knowledge of experimentally observed protein-protein interactions within the host proteome, and associations between short linear peptide motifs and protein domains. The output of the model was a list of host proteins that may interact with specific HIV proteins using specific sites. This list can be used to draft a connectivity map between virus and host, and to determine a set of protein interaction pathways that are significantly enriched by host proteins predicted to be targeted by HIV.

The model was based on the assumption that virus proteins interact with host proteins though a set of conserved linear sequence motifs present in the host proteome. The conserved spatial organization of these motifs on the rapidly evolving HIV proteome supported the assertion that short linear motifs play critical roles in interactions with the host network. The model's predictions led to host protein sets that were crowded by known HIV targeted proteins. This statistically significant enrichment was particularly high along cellular pathways modulated by HIV. The model's predictions were also consistent with experimental data showing phosphorylation events as key targets of HIV when redirecting cell protein networks toward the goal of virus replication.

This study makes two types of predictions of human and virus interactions. The first type of interaction occurs between virus and host proteins. Each of these predictions is supported by testable binding regions on HIV and human proteins. The second type of interaction occurs between the virus and the host KEGG pathways and GO biological processes found to be enriched in host proteins predicted to interact with virus proteins. Both protein and pathway interactions generate hypotheses that can be tested in the lab.

Each predicted virus-host interaction not represented in the experimentally validated set serves as a hypothesis. However, there are so many predictions, and the prediction precision is so low, that it is unreasonable to test all predicted interactions. The value of the predicted virus-host interactions comes when comparing them with other gene level biological interactions, such as the siRNA screens, to formulate hypothesis about specific roles of HIV-human interactions. Specifically, the predicted HIV-human interactions with human proteins that are implicated in an siRNA screen can be tested to see if preventing the interaction has an effect on HIV replication.

In addition to predicting virus-host interactions, this study predicted virus targeted host pathways (Table 2.4). While most of these pathways were already known to be targeted by HIV proteins, there were some that were significantly enriched with predicted virus targeted proteins, while proteins in the validated virus-host interactions showed no enrichment. The cell cycle, Jak-STAT, cytoskeletal regulation, and tight junction KEGG pathways were all significantly enriched in our predictions for some HIV protein, but the corresponding enrichment was not significant for the validated virus-host interactions from NCBI. These pathways offer new hypotheses for cell processes that HIV might need to target.

The methodology applied here for HIV-host protein interactions is applicable to any viruses with multiple sequence alignments and hosts with known interaction networks. Therefore, our approach has potential use in the identification of host

proteins targeted by recently discovered and less studied viruses. The resulting list will be useful for guiding further virus-host interaction experiments, selecting optimal drug therapies, and discovering new antivirus drugs. The systems approach presented here for predicting virus-host protein interactions will benefit from ongoing research on the more specific annotations of short linear motifs and domains involved in protein-protein interactions.

Chapter 3

A bioinformatics approach reveals possible MAPK docking motifs on HIV proteins

3.1 Background

In response to multiple growth factors and cytokines, the mitogen-activated protein kinases (MAPKs) ERK1 and ERK2 play important roles in signal transduction pathways that regulate various cellular processes, which include cell growth, differentiation, gene expression regulation, and cell development [17, 90]. Activation of ERK1 and ERK2 occurs during the G_0/G_1 transition and may be required for progression through the cell cycle [95, 141]. ERK1 and ERK2 are present in all cell types, and are evolutionarily conserved, indicating their importance in cellular signaling pathways [109, 138, 157]. Given that MAPK ERK1 interacts with five HIV proteins [62], and MAPK ERK2 has interactions with ten HIV proteins [90], and both kinases participate in multiple cellular processes, it is likely that it ERK1 and ERK2 are involved in many steps of HIV infection [130].

3.1.1 MAPK and HIV infection

ERK1 and ERK2 have been shown to increase HIV infectivity by phosphorylating a subset of HIV proteins [180]. Inhibiting the phosphorylation of HIV Vif has impaired, but not stopped HIV replication [11, 180]. Prior to HIV replication, the HIV structural protein matrix (MA) must be phosphorylated by ERK2 to allow the HIV pre-integration complex to translocate to the nucleus, where viral replication can proceed [22]. Nef and Tat have been shown to induce the ERK MAPK cascade [144, 167]. ERK1 and ERK2 phosphorylate HIV Nef, Rev, and Tat *in vitro* [180], but the roles of these phosphorylation events in HIV infectivity remain unknown [179]. The inhibition of MAPK phosphorylation has been shown to decrease HIV infectivity, indicating that MA and Vif MAPK-directed phosphorylation events might make good drug targets [22, 180].

3.1.2 MAPK substrate docking

Like all MAPKs, ERK1 and ERK2 phosphorylate their substrates at serine and threonine residues [157]. Before ERK1 and ERK1 can phosphorylate their substrates, they must bind to them at specific docking sites [9]. Two consensus MAPK substrate docking patterns have been proposed for eukaryotes, although some exceptions to these patterns do exist [9, 131]. The Eukaryotic Linear Motif (ELM) Resource [131], a database of peptide motifs that guide protein interactions, has developed patterns that represent the two versions of the MAPK docking site. It refers to these docking sites as LIG_MAPK_1 and LIG_MAPK_2.

The LIG_MAPK_1, or D-site, pattern has two functional regions: a string with two or three basic residues and a chain of alternating hydrophobic residues [9]. One to six residues maintain distance between these regions, helping them interact with distinct regions on MAPKs [10]. The basic component of the motif interacts with MAPKs at a patch of acidic residues, called the common docking (CD) site, while the hydrophobic region of the D-site interacts with a hydrophobic groove close to the CD

site [104]. The LIG_MAPK_2 docking site pattern is simply FXFP, where the letters correspond to amino acids, with X representing any amino acid. The LIG_MAPK_2 docking site is not utilized as a docking site as much as the D-site [131]. Recent work with small-molecule drugs suggests that the D-site could be targeted to disrupt its interaction with MAPKs ERK1 and ERK2. Although using small-molecule drugs to target protein interactions has been difficult in the past [5], there have been recent advances, both experimentally [137] and computationally [126], that could be used to target the docking of MAPK with HIV substrates.

3.1.3 Disrupting protein-protein interactions using small-molecule inhibitors

Most MAPK drugs prevent MAPKs from interacting with ATP by blocking the conserved ATP binding site. The use of the ATP binding site in drug development raises concerns about these drugs' lack of specificity due to similarities among ATP binding sites [123]. New research has suggested that a more specific means of inhibition may be achieved by preventing MAPK substrate docking using small-molecule inhibitors [23, 69]. These protein-protein interaction inhibitors are good drug candidates because they are often cell permeable, and they are more stable than peptide inhibitors [164]. Experimental studies have found small-molecule inhibitors for some protein interactions. For instance, one study found small-molecule inhibitors that stopped apoptosis by blocking the interaction of the Bak BH3 motif with members of the Bcl-2 family [42]. Another study used computational structure modeling and docking to identify small-molecule inhibitors that blocked calcineruin-NFAT signaling by disrupting docking between the calcineruin phosphatase and its substrate [137, 139].

Experimental approaches for identifying small-molecule inhibitors of protein binding are costly and labor intensive [5, 42]. Computationally aided studies like the calcineruin inhibitor development help to reduce experimental drug development by identifying protein interaction sites, and finding small-molecules that will act on

these sites [137]. A pure computational study sought to aid future work concerned with using small-molecule drugs to disrupt protein-protein interactions [126]. By computationally testing all U.S. Food and Drug Administration approved small-molecule drugs for their ability to disrupt protein complexes with known peptide binding sites and available 3D structures, the authors behind the study identified a number of drugs that prevented peptide motif mediated interactions for nuclear receptors and peroxisome components. With knowledge of docking sites on HIV, small-molecule inhibitors might also be developed to disrupt HIV protein phosphorylation by MAPKs, which may hinder HIV replication. However, care must be taken when designing HIV drugs because of strain diversity.

As addressed in Chapter 1, HIV has been classified into different strains, or subtypes, and often these strains combine into recombinant forms [161]. Five subtypes and two recombinants are present in at least 2.5% of the world population, making subtype diversity an issue for drug and vaccine design. HIV subtype has been found to influence transmission and disease progression [161]. The presence and absence of short peptide motifs on HIV proteins has been correlated with patient response to certain therapies [37], indicating that differential docking site usage among strains should be considered when designing drugs.

In this study, we examine MAPKs ERK1/2 docking with HIV proteins from a drug design perspective using multiple alignments of HIV protein sequences taken from different patients, and classified according to subtype. We find that only HIV Nef has docking site pattern hits that cover the majority of protein sequences of the most common HIV subtypes (A1, B, and C). However, the structure of Nef and our *in silico* simulations show that docking at these site is unlikely. Some of the most frequently observed subtypes of HIV proteins MA, Tat, and Vif are missing the docking pattern most often observed in eukaryotic MAPK substrates, whereas HIV Rev does not show the docking pattern on any subtypes. To explain MAPK docking with HIV proteins in a subtype, or strain, independent manner, we impose

Motif	Pattern	Phos(%)	ERK(%)	p-val
Da	[KR]{0,2}[KR].{0,2}[KR].{2,4}[ILVM].[ILVF]	558 (43)	56 (45)	0.299
Db	[KR]{2,3}.{1,6}[ILVM].[ILVF]	513 (39)	51 (41)	0.348
Da U Db	-	620 (48)	69 (56)	0.030
Dc	[KR].{2,6}[ILVM].[ILVF]	841 (65)	92 (75)	0.007
Dd	[KR].{1,3}[KR]{2}	694 (53)	70 (57)	0.231

Table 3.1: Using each of the MAPK docking site patterns, we scanned phosphorylated substrates in the Database of Post Translational Modifications (dbPTM) [97]. We show the number of phosphorylated substrates with pattern matches (Phos column) as well as results for ERK1/2 substrates (ERK column). We used Fisher's exact test to calculate a p-value for the enrichment of pattern hits on ERK1/2 substrates compared to all other phosphorylated proteins. The standard docking site patterns, Da and Db, were not enriched on ERK1/2 substrates, but the union of these patterns, Da U Db, was enriched. Dc, but not Dd, was found to be enriched on ERK1/2 substrates.

slight revisions on the MAPK docking patterns described in the ELM Resource. One such revised motif is present in all major subtypes of HIV proteins known to be phosphorylated by ERK1/2, and is statistically enriched among the substrates of ERK1/2. The use of *in silico* docking indicates the plausibility of the candidate motifs as HIV protein docking sites for ERK1. Our results provide a first step towards identifying the docking site motifs on HIV proteins and await experimental verification.

3.2 Results

3.2.1 Consensus MAPK docking sites on human proteins

As described above, MAPK docking site sequences, found in most eukaryotic MAPK substrates, are presented as two distinct patterns, dubbed LIG_MAPK_1 and LIG_MAPK_2, by the Eukaryotic Linear Motif (ELM) Resource. The LIG_MAPK_2 pattern was not considered in this analysis because it was not found to be enriched in human ERK1/2 substrates, and it was not expressed by the HIV proteome (data not shown). The LIG_MAPK_1, or D-site, pattern has two functional regions: a string with two

or three basic residues and a chain of alternating hydrophobic residues [9]. One to six residues maintain distance between these regions, helping them interact with distinct regions on MAPKs [10]. The ELM Resource describes one version of the D-site (Da), while the current literature contains another frequently observed MAPK docking motif (Db), with a pattern similar, but not identical, to that of Da [9]. Both of the Da and Db motifs have the same biochemical foundations (Table 3.1). The patterns, or regular expressions, of docking motifs Da and Db were constructed to account for MAPK docking sites observed in multiple eukaryotic species [131]. Nonetheless, these motifs can serve as starting templates for the discovery of HIV sequences involved in docking to MAPKs ERK1/2.

To determine the usage of MAPK docking sites in the human proteome, we scanned proteins with documented phosphorylation sites [97] and ERK1/2 substrates [97] with the Da and Db docking site patterns. Since functional peptide motifs tend occur in unstructured regions of proteins [65], tools for motif annotation and discovery do not scan protein regions containing domains [51, 119, 131]. To accomplish this domain filtering, we removed pattern hits falling in Pfam domains [56] in a manner similar to the one used by the ELM Resource [131]. The results presented in Table 3.1 indicated that a combined pattern representing both Da and Db docking motifs, referred to as the Da U Db pattern, was enriched on ERK1/2 substrates relative to all phosphorylated substrates (p-value < 0.03). This statistical enrichment provided evidence supporting the validity of the Da U Db pattern as the MAPK docking site motif among human proteins.

3.2.2 MAPK docking sites on HIV proteins

Since MAPKs ERK1 and ERK2 substrates were statistically enriched with MAPK docking motifs, we hypothesized that the presence of these docking sites on most sequences of HIV proteins MA, Nef, Rev, Tat, and Vif would explain their reported phosphorylation by ERK1 and ERK2. Therefore, we searched for the MAPK docking

A (Da)					B (Db)				
VP	A1	B	C	Total	VP	A1	B	C	Total
MA	6	14	50	34	MA	1	1	1	1
Nef	97	89	98	93	Nef	96	89	91	90
Rev	1	9	1	3	Rev	0	0	1	0
Tat	0	0	0	0	Tat	67	91	44	59
Vif	92	6	97	50	Vif	1	5	61	27
C (Dc)					D (Dd)				
VP	A1	B	C	Total	VP	A1	B	C	Total
MA	95	96	96	96	MA	95	99	97	97
Nef	100	100	100	100	Nef	60	92	86	87
Rev	100	99	96	97	Rev	100	100	100	100
Tat	69	93	45	61	Tat	100	99	100	99
Vif	100	100	100	100	Vif	92	87	100	92

Table 3.2: We searched sequences of HIV proteins using the four MAPK docking site patterns in Table 3.1. Here we present the percentages of HIV subtype sequences with these docking site patterns. The Da and Db patterns were found on the majority of Nef sequences, but they were missing from some subtypes of the other HIV proteins. The Dc pattern occurred on the majority of MA, Nef, Rev, and Vif subtypes. The Dd motif had hits on most sequences of all HIV proteins.

site patterns on HIV sequences gathered from the Los Alamos National Lab (LANL) HIV Sequence Database (http://www.hiv.lanl.gov/), which contains thousands of sequences spanning multiple subtypes and recombinant forms. In this analysis, we considered HIV strains with at least 50 sequences for all HIV proteins known to interact with ERK1/2, leaving three strains to consider: A1, B, and C. Having at least 50 sequences for each strain provided us with enough sequence diversity to assess the conservation of docking pattern hits. Subtypes A1, B, and C are responsible for the majority of the HIV infection around the globe [74], making them appropriate for this study. Figure 3.1 shows the Da and Db motif annotations on multiple sequence alignments of HIV proteins, and Table 3.2 shows the percentages of HIV subtype sequences with docking site matches. The results showed a subtype dependence for the annotations of the Da and Db patterns along HIV proteins.

More than 90% of Nef sequences had the Da motif regardless of subtype, but this motif was absent on most Tat and Rev sequences. Vif subtypes A1 and C, but not

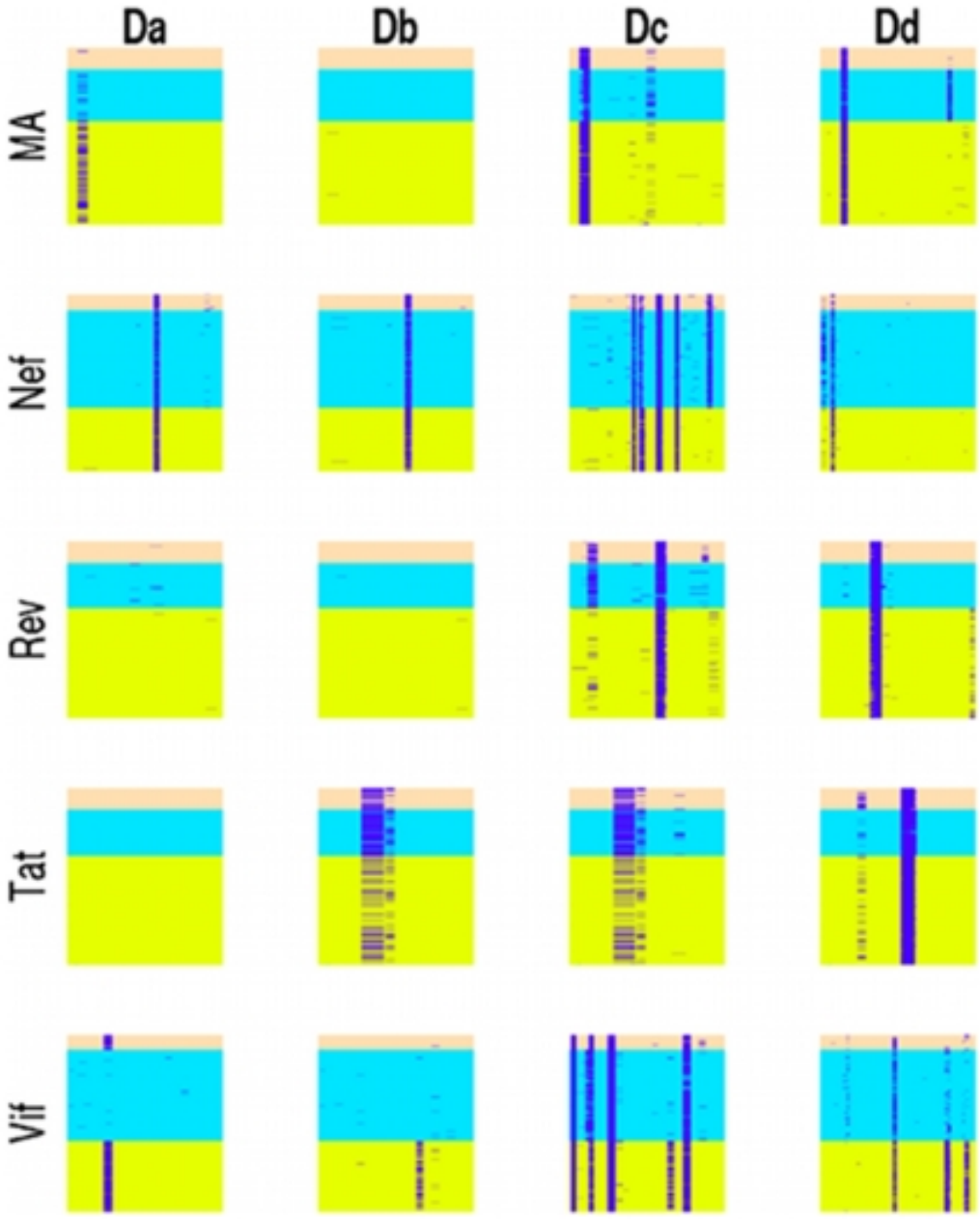


Figure 3.1: Hits for the standard MAPK docking sites, Da and Db, and the proposed MAPK docking sites patterns, Dc and Dd, are annotated in purple on multiple sequence alignments of HIV proteins MA, Nef, Rev, Tat, and Vif. Subtypes in each alignment are represented by different colors: A1 is pink, B is blue, and C is green. Note that motif annotations occur in roughly the same position within a virus subtype.

B, expressed the Da motif. On the other hand, the Db motif was present on Nef, Tat, and subtype C of Vif, but was absent on MA and Rev. The Da and Db motifs occupied different spatial positions along the HIV proteins. The data shown in Figure 3.1 suggested Nef was the only HIV protein for which phosphorylation by ERK1 and ERK2 could be explained by a standard docking site. However, after examining a map of solvent inaccessible residues on Nef [6], we found that the Db motif hit had its last residue in a buried region of the protein. Further investigation using *in silico* docking revealed that both the Nef Da and Db motif hits could not serve as MAPK docking sites (see Methods). Docking at these sites placed the MAPK active site too far from the three possible phosphorylation sites on Nef (Figure 3.2).

Next we revised the Da and Db regular expressions in an attempt to find a motif that would be present on all major subtypes of HIV proteins known to interact with human MAPKs ERK1 and ERK2. We looked at sequences of HIV proteins without a docking motif coincident with the spatial position of the standard motifs. Specifically, we looked at regions along MA and Vif that aligned with the Da motif, as well as regions of Tat and Vif that aligned with the Db motif, but were not annotated with the motif (Figure 3.1). The absence of the Da motif in subtypes A1 and C of MA and subtype B of Vif was caused by a missing basic residue. The absence of Db in some subtypes appeared to be due to mutated hydrophobic residues. Taking cues from these perturbations, we designed a new regular expression for a candidate MAPK docking motif along HIV proteins (Table 3.1), and represented this motif with the symbol Dc. The motif Dc turned out to be present on HIV proteins Nef, Rev, Tat, Vif, and MA in a relatively subtype independent manner (Figure 3.1). We assessed the significance of Dc motif conservation on HIV proteins by comparing it with the conservation found on random protein sequences (see Methods). We found that Dc was significantly conserved on all HIV MAPK substrates compared to random HIV protein sequences (p-value < 0.05).

We also considered whether or not an infrequently observed MAPK docking motif

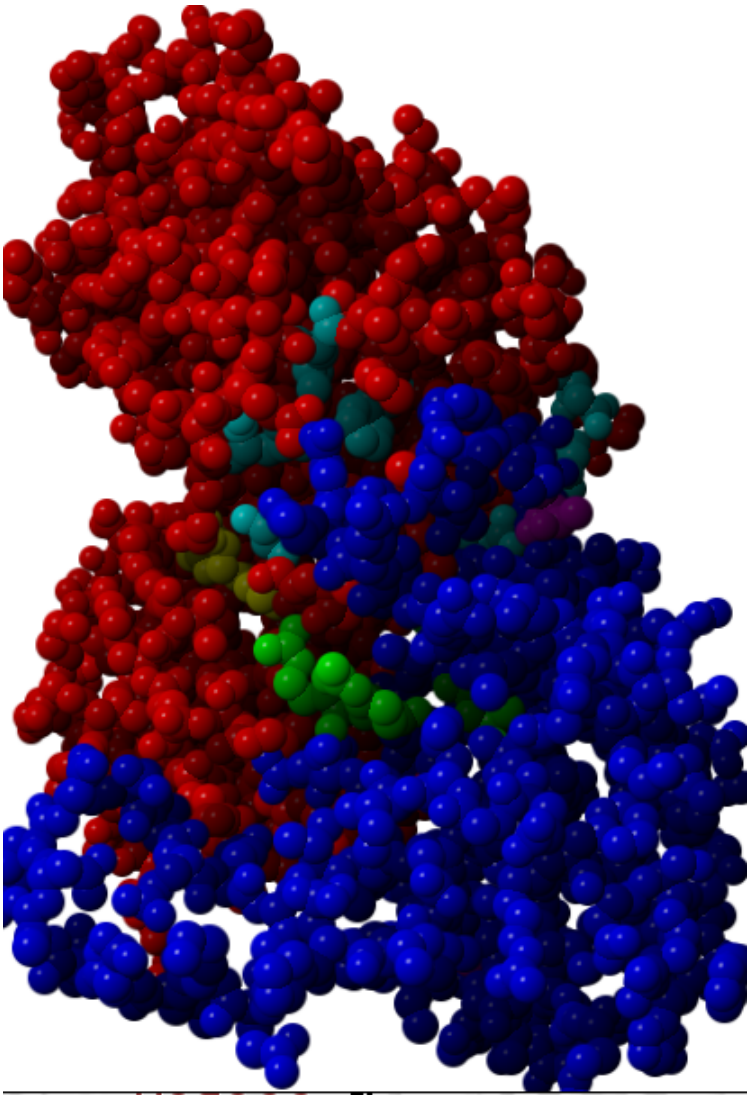


Figure 3.2: *In silico* docking of MAPK ERK1 (red) and HIV Nef (blue) at the MAPK docking groove (cyan) and the standard docking motif hit on Nef (green) did not align the active site of ERK1 (yellow) with any of the three possible Nef phosphorylation site (magenta). This suggested that the standard docking site pattern match on Nef did not function as the real docking site for ERK1.

among human proteins could serve as an HIV subtype-independent MAPK docking site. The docking site pattern for MAPK with thyroid hormone receptor-beta1 (TRβ1) is KGFFRR, where letters represent amino acids. The motif is known to be fully functional, and yet it is missing the hydrophobic portion of the Da and Db motifs [103]. Furthermore, mutational studies showed that only the first and final two basic residues were required for docking, yielding the pattern KXXXRR, where X represents any amino acid [103]. Scanning this motif along the HIV proteome provided new hits, but did not have sufficient coverage along MA, Rev, Tat, and Vif subtypes. We expanded this pattern to include all basic residues in the first and final two amino acids, and allowed variation in the distance between the basic components, resulting in motif Dd, with the regular expression given in Table 3.1. We used this new pattern to scan multiple alignments of HIV proteins, and found hits on the majority of sequences for all HIV proteins known to interact with MAPK (Figure 3.1 and Table 3.2). As with the Dc motif, we found the conservation of the Dd motif on HIV MAPK substrates to be significant (p-value < 0.05) when compared to Dd motif conservation on random protein sequences (see Methods).

The specific sequences matched by MAPK docking motif patterns can be different in human and HIV proteins, and this was best observed by constructing sequence logos from the motif hits on human (Figure 3.3) and HIV (Figure 3.4) proteins known to be phosphorylated by MAPKs ERK1 and ERK2. It was clear from Figures 3.3 and 3.4 that the residue usage for motifs Da, Db, and Dc was similar because all motif hits had basic residues followed by hydrophobic residues. This similar biochemical foundation explained why the candidate docking motif Dc was coincident with Da or Db in the HIV proteome. Our computations based on Fisher's exact test showed the Dc motif to be statistically enriched on ERK1/2 binding partners (Table 3.1). The fact that the Da, Db, Dc motif hits on HIV proteins allow less variation in the spacing residues makes it possible to target these regions with small-molecule drugs while preserving host ERK1/2 activity. The Dd motif had more or less the same

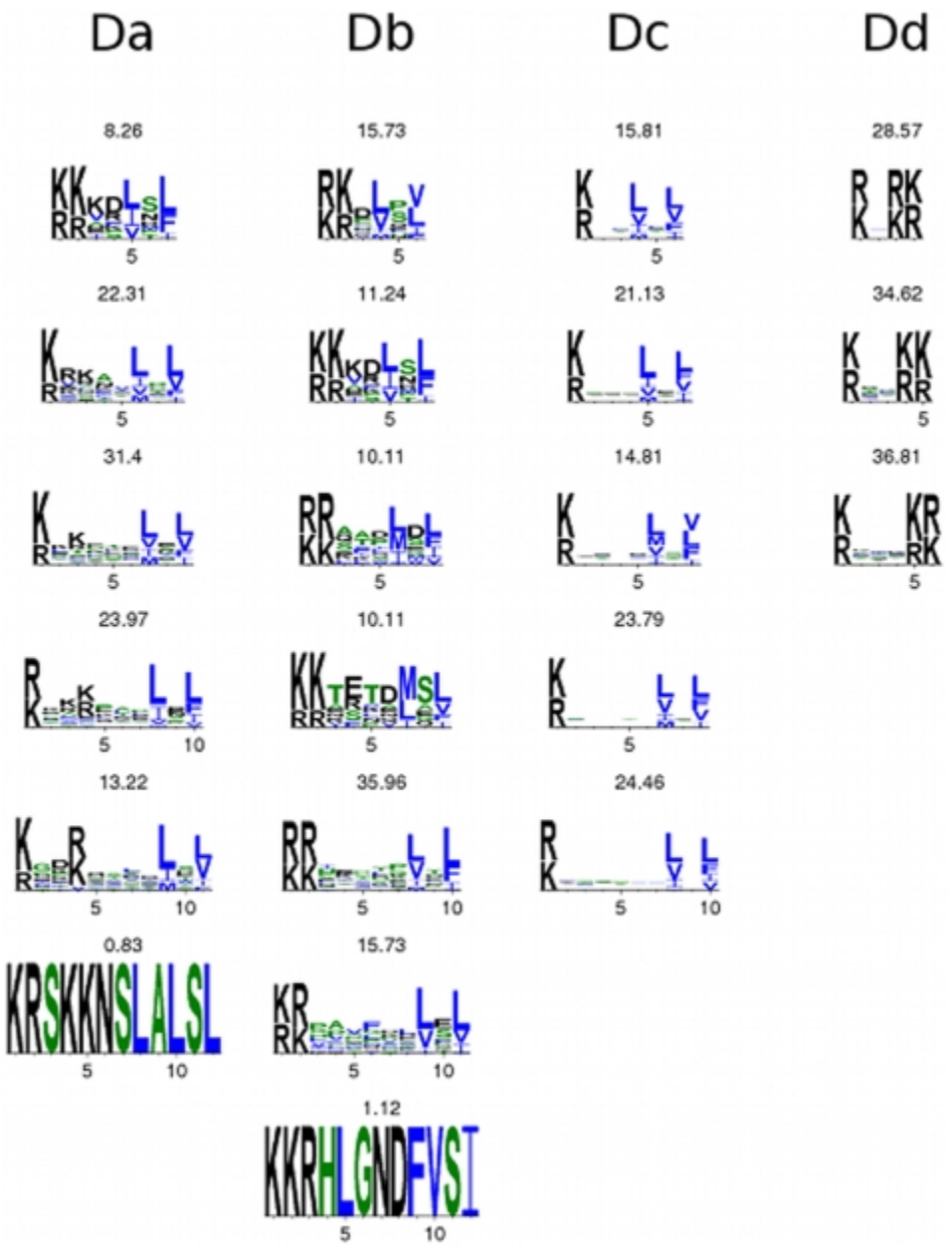


Figure 3.3: For the four MAPK docking sites in the study, we show sequence logos for hits on human proteins. The motifs used here allowed matches with varying lengths. The percentage of motif instances of a certain length is shown above each logo.

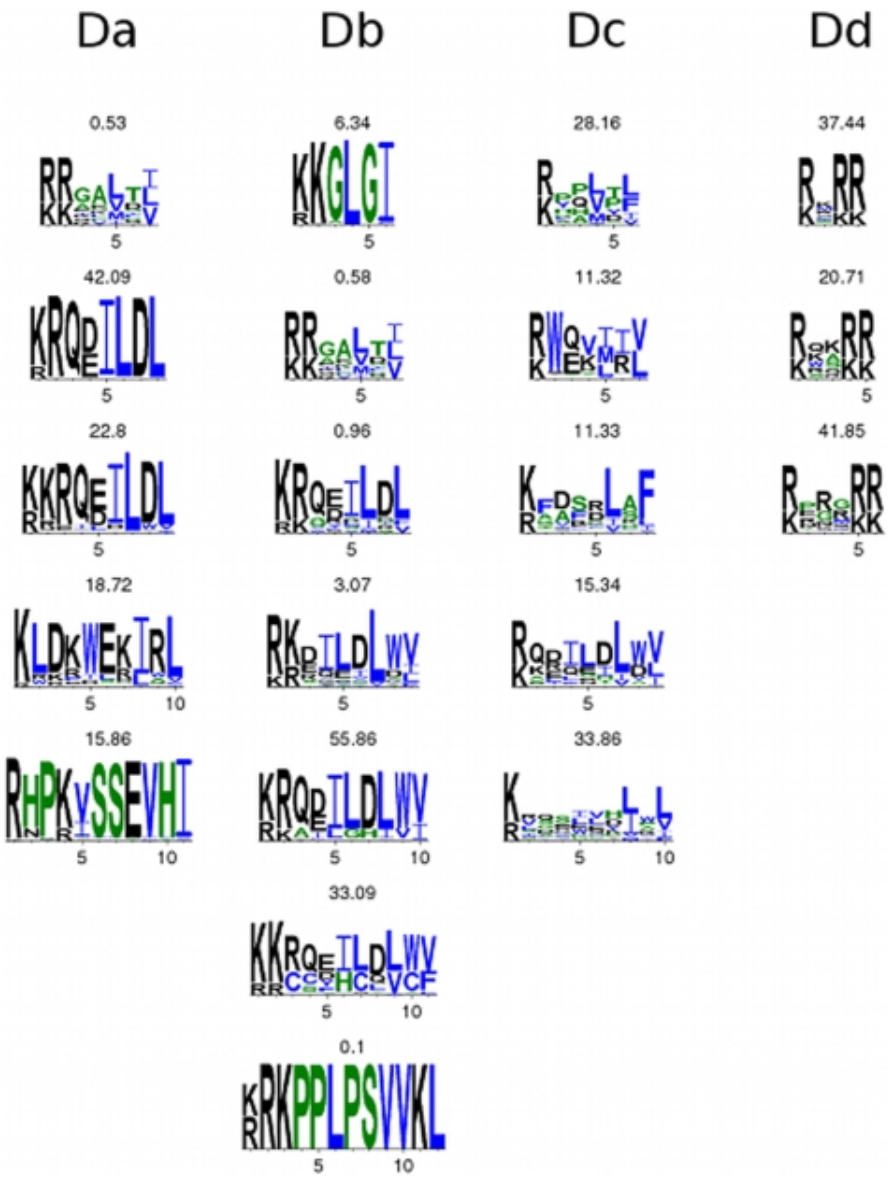


Figure 3.4: Here we show sequence logos for MAPK docking site motif hits on HIV proteins MA, Nef, Rev, Tat, and Vif. For each MAPK docking site motif, we gathered all hits on all HIV proteins and constructed sequence logos for hits with the same length. The percentage of motif instances of a certain length is shown above each logo.

residue usage in human and HIV proteins. This motif is a simple one, and was not statistically enriched among ERK1/2 substrates (Table 3.1).

3.2.3 Candidate docking motifs on HIV Nef

Focusing on ERK1/2 phosphorylation of HIV Nef, we found that neither of the standard Da and Db docking motifs were supported by *in silico* docking (see Methods), and the Db motif failed the solvent accessibility test. This suggested that one of our alternative docking motifs could serve as a potential docking site. Using the solvent accessibility test, we found that none of the Dc pattern hits were likely MAPK docking site candidates, as each one had at least one buried residue. Only the initial N-terminal Dd pattern hit did not overlap with a buried residue. Unfortunately, we were unable to find a Nef protein structure that had both this Dd pattern hit and the proposed Nef phosphorylation sites, so *in silico* docking could not be performed. Testing the functionality of this proposed site awaits further experimentation.

3.2.4 Candidate docking motifs on the HIV protein matrix are supported by structures

In order to further support the feasibility of our new docking patterns as MAPK docking sites, we compared the structures of the HIV matrix protein to those of ERK1/2 substrates. We chose MA in this comparison due to the availability of multiple structures for this protein. Figure 3.5 shows the hierarchical clustering of the HIV MA proteins and ERK1/2 substrates (with known structures) by their pairwise structural similarity, as measured by the TM-score. As explained in the methods section, the TM-score is a normalized measure of structural similarity that ranges between 0 and 1, where a score above 0.20 is considered significant [183]. As expected by their 90% sequence similarity, the HIV MA proteins (shown in bold in Figure 3.5) were clustered together at a high TM-score (0.65). Surprisingly, some of the human

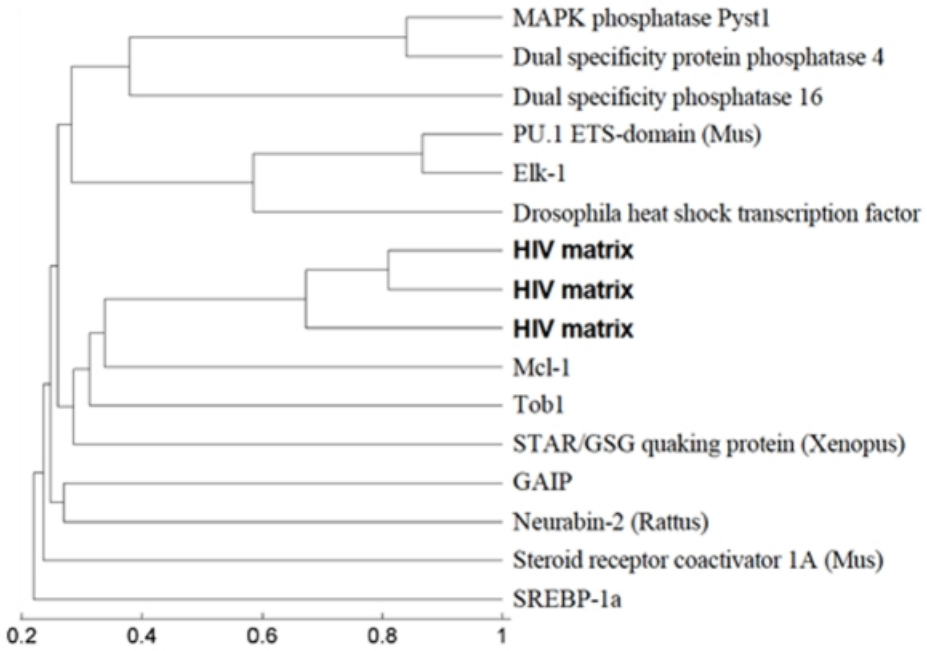


Figure 3.5: Here we present UPGMA clustering of HIV MA proteins and ERK1/2 substrates by pairwise structural alignment TM-scores. All substrates are human proteins unless otherwise indicated. The HIV MA proteins are shown in bold.

ERK1/2 substrates (specifically, Mcl-1, Tob1, and the Xenopus STAR/GSG quaking protein) were found to be structurally more similar to HIV MA proteins than they were to other ERK1/2 substrates. These results were consistent with experimental data showing binding between ERK1/2 and HIV MA.

We next performed *in silico* docking between HIV MA and ERK1 using the ZDOCK server [30]. ZDOCK allows users to force binding between specific residues. The top panel of Figure 3.6 shows ERK1 docked with MA when binding was forced between the hydrophobic portion of Dc and the hydrophobic groove of ERK1. The bottom panel of Figure 3.6 shows ERK1 docked with MA after binding was forced between the Dd motif on MA and the CD site on ERK1 [89]. This figure demonstrates the close proximity of the MA docking site and the ERK1 docking groove after both forced docking interactions. The ATP binding site of ERK1 is bound by the MAPK inhibitor 5-iodotubericidin, colored yellow. Both docking experiments positioned possible phosphorylation sites on MA close to the ATP binding site of ERK1, adding additional evidence that the Dc and Dd patterns on MA could be functional. This was demonstrated in more detail in YASARA (http://www.yasara.org) scenes of the complexes
(www.ncbi.nlm.nih.gov/pmc/articles/PMC2812490/bin/pone.0008942.s001.zip).

3.3 Discussion

In this study we have shown that known MAPK docking motifs occur in a subtype dependent manner on HIV proteins known to interact with human MAPKs ERK1 and ERK2. MAPK substrate docking is known to facilitate phosphorylation. While the detailed role of MAPK phosphorylation in HIV infection has not been established in a clinical setting, ERK1/2 phosphorylation of HIV proteins has been associated with viral infectivity in a number of *in vitro* studies, highlighting the importance of

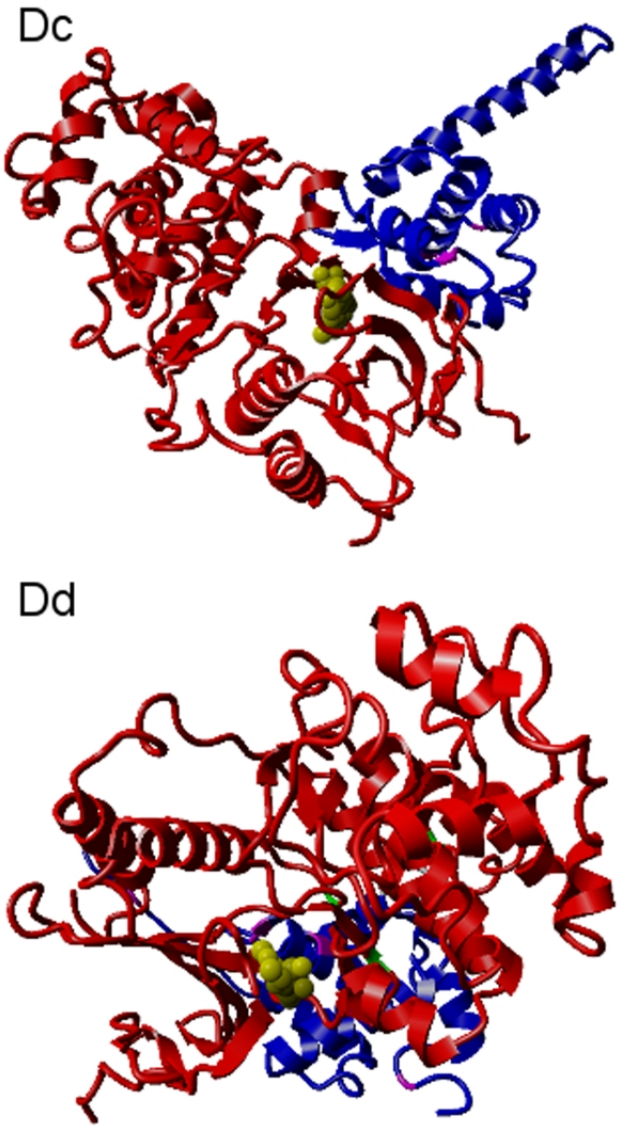


Figure 3.6: *In silico* docking was performed using ZDOCK by forcing ERK1 (red) to dock at the Dc and Dd motifs on HIV MA (blue). In the upper panel, docking was forced to occur between the hydrophobic tail of the Dc motif on MA and the hydrophobic docking groove of ERK1. The lower panel shows the resulting complex when docking was forced between the basic residues of Dd on MA and the CD site of ERK1. The ATP binding site of ERK1 interacts with MAPK inhibitor 5-iodotubericidin (yellow). When the Dd docking site on MA (green) was forced to interact with ERK1, serine phosphorylation sites (magenta) on ERK1 were positioned in close proximity to the ERK1 ATP binding site.

MAPK docking sites in the course of HIV infection. For this reason, we hypothesized that if MAPK phosphorylation of HIV proteins was an essential feature of the progression of HIV infection, then MAPK docking sites along HIV proteins would be subtype independent. This was our motivation for revising known human MAPK docking sites for the case of HIV proteins. Our study suggested two docking motifs, Dc, and Dd, and showed that these appeared on all subtypes of HIV proteins phosphorylated by ERK1/2. These motifs shared biochemical characteristics with motifs used by human proteins that bind to MAPKs. The Dc motif was missing one basic residue form the standard docking motifs, while the Dd motif was missing the hydrophobic portion. The Dd motif had experimental support on human proteins, while the Dc motif did not. *In silico* docking experiments provided evidence supporting the hypothesis that these motifs function as MAPK docking sites along HIV proteins. One of these candidate motifs, Dc, was statistically enriched among the binding partners of ERK1/2. However, the lack of statistical enrichment does not exclude the possibility of the Dd motif being used as a docking site as well.

Current drugs target MAPK kinase activity, but new drugs based on the HIV MAPK docking sites might work better. Existing MAPK drugs, like SB203580, SB202190, and RWJ67657 do not target ERK1/2. FR180204 targets ERK1/2 activity via the ATP binding site, making it susceptible to off target effects. New drugs targeting HIV replication by blocking ERK1/2 phosphorylation of MA and Vif could in theory be incorporated into existing therapy regimens, making it more difficult for an HIV strain to acquire the mutations for resistance to all drugs [41]. By targeting MAPK docking sites, rather than ATP binding sites, drugs can offer more specificity. There is hope that few ERK1/2 substrates will be targeted by drugs specific to HIV docking sequences. Amino acid sequences used by virus and host in the Dc motif were different enough (Figure 3.3 and Figure 3.4) to allow for specific targeting of HIV proteins with drugs. Moreover, the poor structural alignment of HIV MA and other ERK1/2 substrates suggests that HIV specific targeting is possible.

3.4 Conclusion

The standard MAPK docking motifs from the literature could not explain the interactions of MAPKs ERK1 and ERK2 with all subtypes of HIV proteins. The two new motifs we introduced as candidate motifs for ERK1/2 docking were present on subtypes A1, B, and C of HIV proteins known to interact with MAPK. These sites can be tested by mutating key docking site residues on HIV proteins, and observing the effect on their phosphorylation by ERK1 and ERK2. The amino acid composition of the docking motifs on HIV proteins was different enough from the composition found on human ERK1/2 substrates to allow for HIV sequence specific drug targeting using small-molecule drugs. This study can be extended based on a recent computational method for identifying U.S. Food and Drug Administration approved small-molecule drugs that prevent protein-protein interactions [126]. Using our proposed complex of MAPK ERK1 and the HIV matrix protein (MA), approved small-molecule drugs can be screen *in silico* to find those that disrupt the docking between ERK1 and HIV MA. Further annotation of the proposed docking motifs awaits experimental verification.

This study has importance beyond interactions between MAPK and HIV proteins. In Chapter 2, we showed that certain host peptide motifs, like the MAPK docking site discussed here, are found to be conserved on HIV protein sequences taken from different patients. Here we have shown that the transfer of a standard host peptide motif to virus proteins is not as simple as previously thought. We had to make modifications to the docking site provided by the Eukaryotic Linear Motif Resource before we could explain how MAPK could dock with HIV proteins. The requirement for these modifications motivates general questions about host motifs on HIV proteins. What other host peptide motif are missing on HIV proteins due to inadequate patterns? How is the virus utilization of variant host peptide motif patterns beneficial to the virus? These questions help in promoting more studies of virus-host interactions.

3.5 Methods

3.5.1 Human and HIV sequences and motifs

HIV sequence alignments were gathered from the Los Alamos National Laboratory (LANL) HIV Sequence Database, and processed according to [54], but here only subtypes A1, B, and C were used. We gathered 1436 proteins known to be phosphorylated from dbPTM [97], and found 132 of these were phosphorylated by ERK1/2. We scanned all sequences with the four regular expressions for MAPK docking sites. For human proteins, we attempted to rule out false positive hits in a manner similar to the ELM Resource. We removed any pattern hits that overlapped with a Pfam domain [56]. Pfam domains were found for all proteins using the default settings for the stand alone Pfam scan program. Enrichment of docking site pattern hits on ERK1/2 substrates was calculated with a one-tailed Fisher's exact test, using phosphorylated dbPTM substrates as a background set.

3.5.2 Significance of proposed docking site motif conservation on HIV proteins

To assess the significance of our proposed MAPK docking motifs, Dc and Dd, being annotated on most of sequences gathered for HIV proteins MA, Nef, Rev, Tat, and Vif, we devised a control based on randomly constructed HIV protein sequences. Here we describe the control for the HIV matrix protein, but similar steps were used for all HIV proteins phosphorylated by MAPKs ERK1 and ERK2. Using our total set of 987 MA subtype A1, B, and C protein sequences from the LANL database, we estimated the probability of amino acid occurrence as well as amino acid transition probabilities, i.e. the probability of seeing amino acid β follow amino acid α in MA protein sequences. We constructed one random MA protein sequence for every real MA protein sequence by first sampling an initial amino acid based on the single

amino acid probabilities, and then using the amino acid transition probabilities to sample subsequent amino acids and build the rest of the random protein sequence until it was as long as the real one.

We made one hundred sets of random MA protein sequences, each containing 987 random MA sequences, and matched all proteins in each set against patterns for the Dc and Dd docking sites. For each random protein set, we calculated the conservation of the proposed docking sites across random MA protein sequences in the set, and compared this conservation to the docking site conservation observed for real MA protein sequences from the LANL database. To obtain a p-value for the conservation of Dc and Dd on real MA protein sequences, we recorded the number of random sets where the conservation of the proposed docking site was equal to or greater than the conservation observed for real MA protein sequences. We found that both Dc and Dd docking site motifs were significantly conserved compared to random protein sequences (p-value < 0.05), i.e. the proposed docking motifs had higher conservation on real MA protein sequences than on random protein sequences in more than 95 of the random sequence sets. Using a similar control for HIV Nef, Rev, Tat, and Vif proteins, we found Dc and Dd motifs to be significantly conserved for all HIV MAPK substrates, which made it more likely that they were guiding interactions with MAPKs ERK1 and ERK2.

3.5.3 Structure analysis

We performed a BLAST [4] search on the Protein Data Bank (PDB) [15] to identify known ERK1/2 substrate protein structures (E-value threshold of 1e-10). We collected the proteins that had less than 150 residues, to be comparable to HIV MA in size, and only kept the top hit in each BLAST result set. A pairwise structural alignment of each of three MA structures [PDB: 1hiw, PDB: 1uph, PDB: 2hmx] against each MAPK substrate structure was performed using Vorometric [143]. The alignments were filtered by a TM-score [183] of 0.25, resulting in 10 ERK1/2 substrate

structures. The PDB identifiers for the 10 substrates are as follows: PDB: 1mkpA - MAPK phosphatase Pyst1, PDB: 3ezzA - Dual specificity protein phosphatase 4, PDB: 2vswA - Dual specificity phosphatase 16, PDB: 1pueE - PU.1 ETS-domain, PDB: 1duxC - Elk-1, PDB: 1hksA - Drosophila heat shock transcription factor, PDB: 2pqkA - Mcl-1, PDB: 2z15A - Tob1, PDB: 2bl5A - STAR/GSG quaking protein, PDB: 1cmzA - GAIP, PDB: 2g5mB - Neurabin-2, PDB: 1oj5A - Steroid receptor coactivator 1A, PDB: 1am9A - SREBP-1a.

3.5.4 Docking

Docking in Figure 3.2 was performed with the ZDOCK server (http://zdock.bu.edu) to find the most likely complex of ERK1 [PDB: 2zoqA] and HIV Nef [PDB: 2nef] when binding was forced between the proposed docking site on Nef (Arg105, Arg106, Leu110, and Leu112) and the CD site of ERK1 (Glu98, Asp179, Asp335, and Asp338 [89]).

For the top panel of Figure 3.6, ZDOCK was also used to calculate the most probable complex of ERK1 [PDB: 2zoqA] and HIV MA [PDB: 1uphA] when binding was forced between the hydrophobic tail residues of the Dc motif on MA (Ile19 and Leu21) and the hydrophobic docking groove of ERK1 (Thr127, Leu132, Leu138, and Phe146 [89]).

For the bottom panel of Figure 3.6, the ZDOCK server was used to calculate the most probable complex of ERK1 [PDB: 2zoqA] and HIV MA [PDB: 1uphA] when binding was forced between the basic residues of the Dd motif on MA (Arg22, Lys26, and Lys27) and the CD site of ERK1.

Chapter 4

Modularity in protein interaction network hubs predicts viral host-pathogen interactions

4.1 Background

As noted in Chapter 1, protein-protein interactions from single organisms have been organized into networks, where proteins are nodes and network edges represent interactions between proteins [67]. These protein-protein interaction (PPI) networks have two types of nodes, hubs and non-hubs [8]. Characterizing protein nodes in this manner happened after PPI networks were observed to be scale-free, i.e. the distribution of the number of proteins each network protein interacts with follows a power law where a small percentage of network proteins have the majority of interactions in the network [80]. These highly connected network proteins are referred to as network hubs. Although the scale-freeness of networks has been debated [68, 132, 156], the property indicates that networks are robust to perturbations on random nodes because most nodes are not hubs [3, 101], and the specific removal of hub nodes can drastically alter network structure by removing many interactions [3, 101].

The study of host network hubs in virus-host systems biology is important for two reasons. First, HCV, HIV, influenza virus, Epstein-Barr virus, and other human viruses preferentially interact with hubs in the human PPI network [26, 40, 50, 160]. It has been speculated that this viral targeting of host hubs is an efficient way to rewire the host network [26]. Second, host hubs are important for the study of virus interactions with host networks because they have been well studied, and have a number of important properties that could explain virus-host interactions. Knowledge of host hub properties from network systems biology provides a number of host protein features that might be targeted by viruses.

Most of what has been learned about host hubs is useful to consider when looking for properties of virus targeted host proteins. Hubs are evolutionarily conserved [61]. When compared across organisms, hubs have lower mutation rates than other network proteins [14]. Furthermore, comparing the number of interactions for hub proteins across human, fly, worm, and yeast revealed that hub proteins have similar numbers of interactions across all organisms [57]. Studies of high quality PPI networks have revealed a correlation between the number of PPIs a protein participates in and its importance to the cell [73]. This has led to the conclusion that hubs are essential for cell survival [14, 67]. Hubs also have only small changes in gene expression across different conditions as compared to other network proteins [106]. Proteins mutated in cancer are more likely to be network hubs [81]. Finally, hubs allow for network evolution. Gene duplications are observed more often for hub proteins, creating redundancy in the PPI network that can lead to neofunctionalization [83, 163].

4.1.1 Intermodular and intramodular hubs

Since the study of hubs is important to the study of networks, or interactomes, hubs have been investigated further, and it has been revealed that they can be divided into two classes, or modes, by examining their co-expression with their interactome

neighbors [162]. In human, hubs are classified as intermodular or intramodular. Intermodular hubs are defined as being co-expressed with their neighbors in certain tissues, while intramodular hubs are characterized by co-expression with their neighbors in most tissues [162]. This hub distinction based on gene co-expression is important for cancer. It has been demonstrated that the change between intermodular and intramodular modes is predictive of breast cancer patient survival [162].

The case for hub modularity was first made in yeast, where hubs were classified as 'date' or 'party' using time series expression data to look at the co-expression of hubs and the proteins they were observed to bind to in the yeast interactome [67]. The distribution of hub interactome neighbor co-expression values was observed to have two modes. Hubs in the party hub mode were described as having a higher level of interactome neighbor co-expression than hubs in the date hub mode. Non-hub proteins did not show this bimodal distribution of hub and neighbor co-expression. Further analysis revealed that date hubs served as connections between protein modules and complexes, while party hubs acted as their central components [59]. This observation led to the re-branding of date and party hubs as intermodular and intramodular hubs, respectively. Taylor et al. extended these terms to include human hubs, demonstrating that intermodular hubs were co-expressed with their neighbors in specific tissues, while intramodular hubs showed neighbor co-expression across most tissues [162]. These hub classifications have been debated for both yeast and human. In yeast, the analysis of larger interactome datasets failed to replicate the hub distinction [12, 13]. In human, using a more controlled normalization of the expression data used by Taylor et al. caused the hub class distinction to disappear [1]. Specifically, the GCRMA algorithm [176], which controls for probe affinity, was used instead of the Affymetrix MAS5 algorithm [77] adopted by Taylor et al..

Intermodular and intramodular hubs have specific properties in both yeast and human [59, 162]. Intermodular hubs provide temporally and spatially constricted connections between intramodular hubs, which serve as components of cell machinery

[59]. These connections allow intermodular hubs to direct macromolecular complexes in a time and space dependent manner. Intermodular hub proteins have been found to be larger than intramodular hub proteins, and have more unique domains and peptide binding motifs than intramodular hubs [162]. Intramodular hubs have been found to share more molecular functions with their PPI neighbors than intermodular hubs [162]. As expected of hubs that modulate cellular activity, intermodular hubs were found to be enriched in cell signaling domains [162]. In yeast, intermodular hubs are more disordered, or less structured, than intermodular hubs [52, 152]. This is consistent with the observation that intermodular hubs have few binding surfaces, while intramodular hubs have multiple, similar binding regions that allow many interactions to occur at once [88]. Note that this is not inconsistent with the finding that human intermodular hubs have more unique binding regions than intramodular hubs. In fact, human intramodular hubs were found to have greater globularity, or domain coverage, than intermodular hubs [162], which is consistent with the yeast hub result that intramodular hubs have more protein binding regions. Intramodular hubs have also been found to have lower evolutionary rates than intermodular hubs [60]. This may be caused by the greater globularity and more structured regions in intramodular hubs as compared to intermodular ones [84].

In this chapter we provide new evidence for the inter/intramodular hub distinction, and ask if hub proteins that interact with virus proteins favor one hub type over the other. The work presented here is important for two reasons. First, it contributes to the network biology field by reaffirming the existence of two hub classes by showing that viruses prefer one hub class over another. Second, it aids the study of virus-host networks by refining the observation that viruses preferentially interact with hub proteins. This refinement focuses the multiple hub properties that might be behind the virus preference into a few testable hypotheses for future studies.

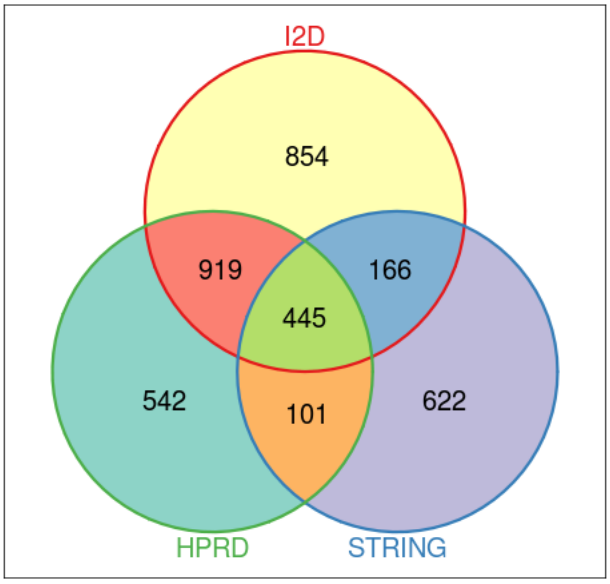


Figure 4.1: In this study we used three human interaction networks, I2D, STRING, and HPRD. To find hubs for each network, we took the top 20% most connected nodes. Each network had a different hub set due to differing network size and connectivity. Here we show the overlap between hub sets for the three networks.

4.2 Results

4.2.1 Hub classification

In this study, we re-examined the hub class hypothesis in human with the goal of investigating virus hub preference. We adopted the same approach used by Taylor et al., but substituted their expression dataset with one from COXPRESdb [122], which is larger in terms of genes and samples. To ensure that our results were robust to interaction network changes, we used three separate human interaction networks from the Interologous Interaction Database (I2D) [21], the Search Tool for the Retrieval of Interacting Genes/proteins (STRING) [174], and the Human Protein Reaction Database (HPRD) [129]. I2D, with roughly 150 thousand edges connecting 11.5 thousand proteins, is the largest network, but it also contains the most interaction

predictions. HPRD has around 35 thousand edges between nine thousand proteins. This network has high quality interactions curated from the literature. Like I2D, STRING has many predicted interactions to complement interactions from the literature, resulting in a network of about 140 thousand interactions between six thousand proteins. For each network, we designated the 20% most connected proteins as hubs [16]. Due to connectivity differences, each interactome resulted in a different set of hubs (Figure 4.1). To make the hub classifications for each network, we used Pearson correlation coefficients (PCCs) from COXPRESdb to indicate correlated gene expression between hub proteins and their neighbors, and obtained an average PCC for each hub by averaging the PCCs of the hub's neighbors [67, 162]. We refer to the distribution of average PCC values for hubs as dPCC. For each interactome's dPCC, we used a likelihood ratio test [53] to confirm that a bimodal distribution was a significantly better fit for the data than a unimodal distribution. The resulting bimodal distributions for each network did not allow an easy separation between inter/intramodular hubs because of the large overlap between the two modes, so we fit a mixture of two Gaussian distributions to binned versions of each network's dPCC (Figure 4.2a). We found that limiting hubs in one network to those present in one or both of the other two networks resulted in a stronger bimodal signal for each dPCC (Figure 4.2b). For the remainder of the study, we limited hubs in one network to only proteins classified as hubs in one of the other two networks.

We obtained intermodular and intramodular hub sets for each interactome by assigning hubs to the most likely mode of the interactome's dPCC. As defined before, intramodular hubs in the higher mode were co-expressed with their neighbors in most tissues, while intermodular hubs in the lower mode were co-expressed with their neighbors in certain tissues. Like the hub, non-hub classification, interactomes differed in their inter/intramodular hub sets, but there was much overlap despite the networks having different connectivities (Figure 4.2c). Final hub classifications were determined by the most frequently observed hub class across the three networks.

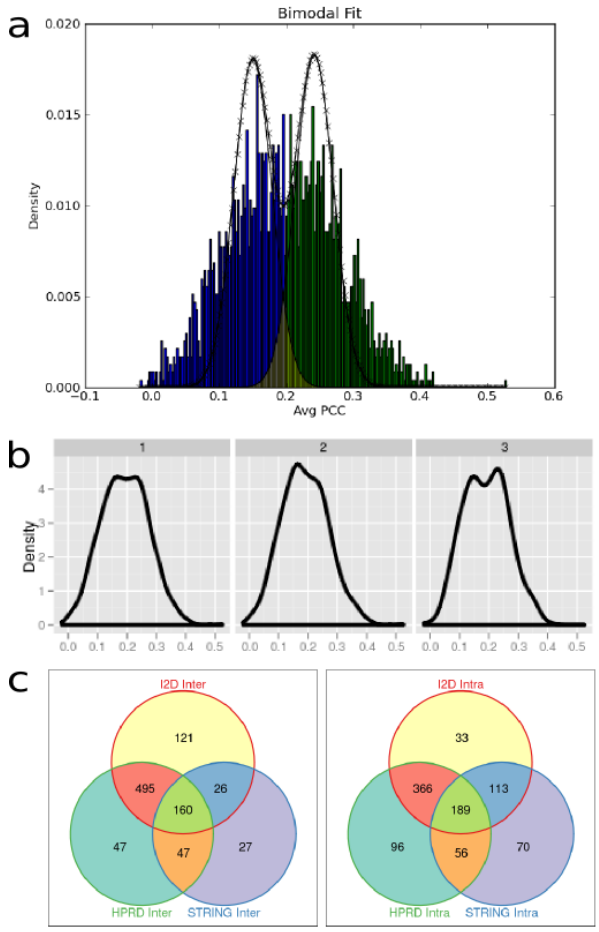


Figure 4.2: This figure illustrates how we have classified intermodular and intramodular hubs. (a) The distribution of hub average PCC values (dPCC) for I2D hubs was fit with a bimodal distribution. Hubs were assigned to either the intermodular mode (blue) or intramodular mode (green) of the bimodal distribution. (b) The dPCC each for network hubs became more bimodal as the number of interactomes the hub had to be present in increased. The density curves shown here for the I2D interactome are for all hubs present in I2D (1), hubs in I2D and STRING or HPRD (2), and hubs present in all three interactomes (3). (c) Inter/intramodular hub classifications were compared for I2D, STRING, and HPRD. In an effort to make are results robust to network changes, we focused on hubs with the same classification in at least two networks.

Hubs that were only present in two interactomes, and had conflicting class assignments, were not considered in the study. Hub classifications were stochastic because of equally likely hub assignments, making hub class assignments differ slightly between runs, but our results were qualitatively similar for all runs. Here we present the results for one run.

4.2.2 Hub class properties

Using hubs with consistent inter/intramodular classifications in at least two networks, we examined biological features of intermodular and intramodular hubs. As reported by Taylor et al., intermodular hubs had more linear motifs [131] per residue and more unique SMART domains [99] per protein than intramodular hubs (one-tailed Wilcoxon tests, p-value < 0.03 for both). Taylor et al. also reported that intramodular hubs were longer than intermodular hubs, but we found no difference in protein length between the hub classes. We used DAVID [43] to find KEGG pathways [85] and Gene Ontology [7] terms enriched in each hub class compared to all hubs. Intermodular hubs were enriched in genes involved in signal transduction, kinase cascades, anatomical structure morphogenesis and development, cellular development, and multicellular organismal development (Bonferroni-corrected p-value $<$ 0.01). Intramodular hubs were enriched in genes annotated with translation, mRNA metabolic process, RNA splicing, ribosome and proteasome components, nucleotide binding, pyrophosphatase activity, and cell cycle (Bonferroni-corrected p-value $<$ 0.01). These hub enriched terms are consistent with what has been reported for hubs in both yeast and human [59, 162].

4.2.3 Virus hub preference

With our hub classes established, we gathered three sets of virus-host interactions to test for an association between hub class and hub proteins that interact with virus

Average hub-neighbor co-expression			
Virus	Intermodular	Intramodular	P-value
HCV	73/655	56/668	0.948350202594
HIV	93/635	130/594	0.00380164624435
Influenza	40/688	88/636	4.77173152977e-06
Median hub-neighbor co-expression			
Virus	Intermodular	Intramodular	P-value
HCV	71/646	52/646	0.958559296013
HIV	95/622	121/577	0.01952745875
Influenza	39/678	84/614	6.9130570599e-06

Table 4.1: The table shows the number of virus targeted inter/intramodular hubs considered when testing for an association between hub class and virus interaction, as well as the results from a one-tailed Fisher's exact test. HIV and the influenza virus target proteins that are enriched in intramodular hubs, while HCV targeted proteins almost show a preference for intermodular hubs. Calculations were performed using the average and median to calculate a measure of co-expression between hubs and their interacting neighbors (see text).

proteins. We found human proteins that interacted with an HIV protein by combining VirusMINT [29] and the NCBI HIV-1, Human Protein Interaction Database [62]. For HCV, we relied on a collection of virus-host interactions identified by yeast two-hybrid screens and a literature search [40]. For influenza virus, we used curated interactions gathered to study influenza virus replication [91].

In an attempt to detect a trend for virus-host interactions, we separately tested hub sets targeted by HCV, HIV, and influenza virus for an inter/intramodular hub class preference using a one-tailed Fisher's exact test (Table 4.1). HIV and influenza targeted hubs tended to be intramodular (Figure 4.3). On the other hand, HCV targeted hubs were almost significantly associated with intermodular hubs (Fisher's exact test, p-value < 0.055).

To further study the viruses, we turned to small interfering RNA (siRNA) screens for virus dependency factors (VDFs) [24, 91, 102, 181]. In these screens, host genes were knocked down in virus infected cells, and the effect on the virus is observed. Genes whose expression depletion negatively affected the virus were recorded as VDF hits, i.e. host factors that are required by the virus [24]. For HIV, we merged

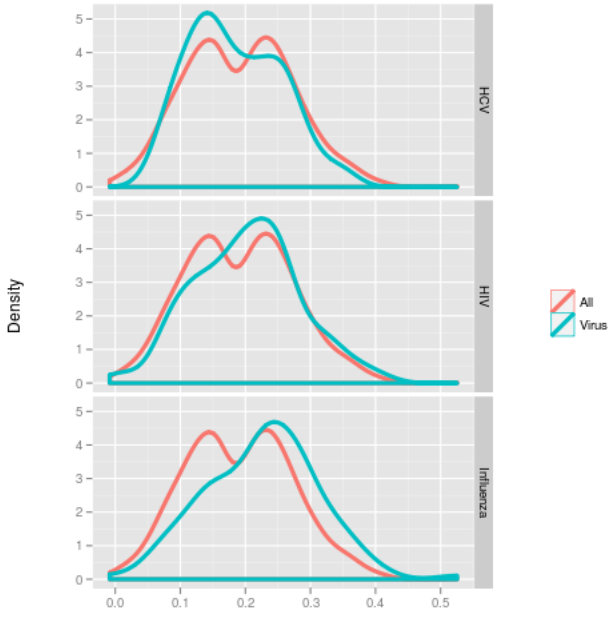


Figure 4.3: This figure visually compares the virus hub preference for HCV, HIV, and influenza virus. For each virus, dPCC density curves for virus targeted hubs were plotted against the dPCC density curve for all intermodular and intramodular hubs in the human I2D network. We used only hubs present in I2D and another human network. Hubs that interacted with HIV or influenza virus preferred the intramodular mode of the dPCC bimodal distribution.

results from four siRNA screens that searched for VDFs required for HIV replication [24, 181]. These datasets of had few genes in common, but genes across datasets often had roles in the same pathways [181]. HCV siRNA results were taken from several screens that covered the full HCV life cycle [102]. Influenza virus screen results were taken from a study of VDFs involved in influenza replication [91]. Just as host proteins found to bind to virus proteins were enriched in hubs, we found that VDFs had more interactome neighbors than genes with no virus association (one-sided Wilcoxon test, HCV: p-value $<$ 6e-8, HIV: p-value $<$ 3e-11, influenza: p-value $<$ 9e-19). Using Fisher's exact test again, we found that HIV still showed an intramodular hub preference for the siRNA data, while HCV and the influenza virus did not (Table 4.2).

The bimodality of each network's dPCC might have been caused by a significant difference between hub classes in terms of the number of interacting neighbors, or degree, considered for each hub, i.e. proteins in one hub class would have significantly different numbers of interaction neighbors than proteins in the other hub class [162]. Taylor et al. addressed this concern and found no difference in the degree distributions of the hub classes [162]. We compared intermodular and intramodular hub degree distributions for each network separately using a one-tailed Wilcoxon test. For the I2D and STRING interactomes, we found that intramodular hubs had more neighbors than intermodular hubs (p-value $<$ 3.5e-12 and p-value $<$ 5.5e-12, respectively). For HPRD, neither hub class had more interactions. Since we found a degree distribution difference between inter/intramodular hubs for the I2D and STRING networks, we repeated the classification of host hub proteins using the median instead of the average when summarizing hub neighbor co-expression. If the degree difference between inter/intramodular hubs was causing the hub distinction, using the median instead of the average would solve this problem by removing the influence of outliers in a hub's collection of neighbor co-expression correlations. Using the median instead of the average did not produce qualitatively different results

Average hub-neighbor co-expression			
Virus	Intermodular	Intramodular	P-value
HCV	25/703	38/686	0.0581221952106
HIV	58/670	106/618	3.79707703564e-05
Influenza	34/694	36/688	0.441928925248
Median hub-neighbor co-expression			
Virus	Intermodular	Intramodular	P-value
HCV	21/696	41/657	0.00477984148923
HIV	56/661	106/592	8.58550353701e-06
Influenza	36/681	34/664	0.599420611669

Table 4.2: The table shows the number of inter/intramodular hubs considered when testing for an association between hub class and virus dependency factors, as well as the results from a one-tailed Fisher's exact test. Only host factors required for HIV replication are associated with intramodular hubs. Calculations were performed using the average and median to calculate a measure of co-expression between hubs and their interacting neighbors (see text).

for virus-host protein interactions (Table 4.1). For the siRNA data, the association between HCV host dependency factors and intramodular hubs became significant (Table 4.2).

For each virus, we intersected siRNA screen hits with host proteins that interacted with virus proteins to arrive at virus-host protein binding interactions that might play an important role in the virus life cycle. Figure 4.4 shows the network of connections between of these virus targeted inter/intramodular hubs and virus proteins. Human hubs were placed into functional categories according to the literature [24, 162]. HIV targeted intramodular hubs included proteins involved with transcription and splicing, nuclear transport, HIV cell entry and budding, and the proteasome. The four intramodular HCV targeted hubs were serine/threonine-protein kinase TBK1, actin-modulating protein CFL1, nuclear import protein IPO5, and CDK6. Differing from HCV and HIV targeted intramodular hubs, some influenza virus targeted intramodular hubs were involved in apoptosis and the cell cycle. Intermodular HCV targeted hubs included proteins involved in transcription, translation, and cellular transport. Intermodular hubs found to interact with all three viruses

included members of the Jak/STAT and MAPK signaling pathways.

4.3 Discussion

To find a biological reason for the hub preference observed for HIV and influenza virus, we focused on the roles of human protein complexes in virus life cycles. HCV, HIV, and influenza virus all encode less than 20 proteins, and must rely on human proteins to accomplish virus replication [102, 160]. There is much anecdotal evidence that HIV proteins interact with human complexes to accomplish important roles in HIV replication [24]. HIV proteins interact with the mediator, P-TEFb, and elongin complexes to accomplish HIV transcription. HIV splicing is directed by HIV proteins interacting with hnRNP complexes. Interactions of HIV proteins with the chaperone containing TCP1 complex might play a role in HIV budding. For HCV infection, the role of human complexes is less clear. An siRNA screen indicated that the Golgi-associated retrograde transport complex might play a role in HCV replication [102]. Influenza virus has been predicted to interact with host complexes such as the ribosome, proteasome, and splicesome [91]. To establish a link between human protein complexes and viruses, we counted the number of complexes a human protein participated in using complexes listed in HPRD. In our hub set for HPRD, more than half of the hubs participated in at least one complex, which was significantly more than the fraction of complexed proteins found in the entire network (p-value $<$ 9e-57). For each of the three viruses, we used a one-tailed Wilcoxon test to show that virus targeted human proteins participated in more protein complexes than other human proteins (HIV: p-value $<$ 0.002, HCV: p-value $<$ 0.02, influenza: p-value $<$ 6e-4). This is expected from the observation that host hub proteins participate in more complexes than other network proteins, so we conducted a further test with HIV that revealed that virus targeted hub proteins participated in more complexes than other hub proteins (one-sided Fisher's exact test, p-value $<$ 0.03).

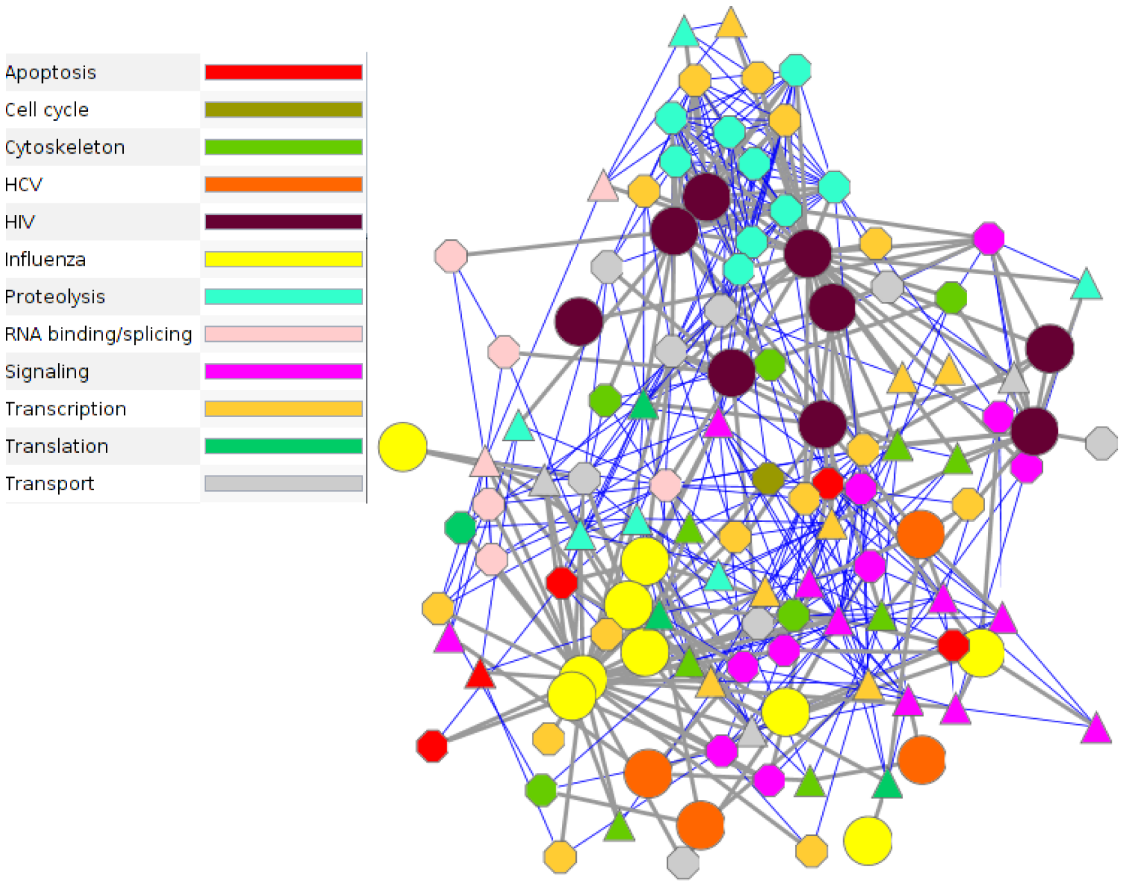


Figure 4.4: Here we show the inter/intramodular hubs targeted by HCV, HIV, and the influenza virus. Virus proteins are large circles. Intermodular hubs are triangles, and intramodular hubs are octagons. Virus-host interactions are thick lines, while human-human interactions are thin and blue. Human hubs are colored by biological function.

In yeast, intermodular hubs connect modules and complexes, while intramodular hubs serve as their central components [59]. Using protein complexes described in HPRD, we confirmed that human intramodular hubs participated in more complexes than intermodular hubs (one-sided Wilcoxon test, p-value < 0.02). If viruses must modulate the activity of protein complexes, they could do so by interacting with complexes directly through intramodular hubs, or they could interact with proteins that form connections between complexes, the intermodular hubs. Given that HIV and influenza virus prefer to interact with human proteins involved in many complexes, and intramodular hubs are involved in more complexes than intermodular hubs, we propose that some virus proteins interact preferentially with intramodular hubs because they are involved in more protein complexes than intermodular hubs.

4.4 Conclusion

Our results have further demonstrated the distinction between intermodular and intramodular hubs. We found that by considering hubs present in multiple networks, this distinction becomes more evident. We used this observation to classify human hubs that have direct protein interactions with virus proteins, and found that HIV and influenza virus targeted hubs were more likely to be intramodular than intermodular. HCV targeted proteins did not show a significant hub type preference, but since the siRNA data did show an intramodular preference, this may change as more interactions are gathered. Compared to intramodular hubs, intermodular hubs had more linear motifs per residue and more unique SMART domains. Intermodular hubs were enriched in signaling and developmental pathways while intramodular hubs were enriched in translation and mRNA processing. Intramodular hubs participated in more protein complexes than intermodular hubs, and, given that HIV and influenza virus proteins prefer to interact with members of many complexes, this bias might be causing the hub class preference.

Knowing that some viruses have an intramodular hub preference aids the study of virus proteins in two ways. First, a virus intramodular hub preference is beneficial for virus-host network studies that are functionally annotating virus proteins based on their host binding partners. In single organisms networks, proteins that interact with each other are often in the same cellular pathways, and share the same functions [145]. This observation has been used to annotate proteins of unknown function with the functions of proteins with which they interact [86, 148]. This annotation method has been extended to virus-host networks, annotating virus proteins based on the functions of their interacting host proteins [26]. Since intramodular hubs are more likely to be functionally similar to the proteins they interact with [162], the virus intramodular hub preference justifies annotating virus proteins using virus-host networks, while an intermodular hub preference would cast doubt on this annotation method.

The second way that a virus intramodular hub preference aids the study of virus proteins is by providing protein features that viruses might be targeting when they interact with host proteins. Here we provided evidence that virus proteins prefer to interact with intramodular hubs because they participate in host protein complexes more often than intermodular hubs. Intramodular hub proteins are also more structured than intermodular hubs, participate in more cellular housekeeping activities than intermodular hubs, and evolve at lower rates than intermodular hubs. Further study is needed to see if these features might also be targeted by virus proteins, and to determine the importance of each feature to viruses.

4.5 Methods

4.5.1 Human interaction networks

We converted the human I2D (accessed October 2009), human STRING (version 8.2), and HPRD (release 8) interactomes to networks of Entrez Gene IDs using

ID mapping provided by UniProt [34], gProfiler [135], and NCBI's eutils. Network hubs were found for each interactome separately by locating the lowest degree in the top 20% of connected genes, and taking all genes with at least this degree. Bimodal curves for each interactome's hub average PCC distribution were found using gradient descent to minimize the log-likelihood of a binned distribution. The number of bins for each distribution was determined by dividing the total number of data points by a bin divisor. For individual networks, we used a bin divisor of 11 for I2D and 10 for STRING and HPRD. For distributions made using hubs present in at least two networks, we fit a bimodal curve using a bin divisor of 6 for STRING and 9 for I2D and HPRD.

4.5.2 Virus-host interaction networks

We gathered HIV-human protein interactions from NCBI and VirusMINT. Both virus-host interaction datasets label interactions by type. We focused on interaction types describing direct protein interactions, or modifications. We used VirusMINT interactions labeled with MINT interaction IDs 0006 (anti bait coimmunoprecipitation), 0007 (anti tag coimmunoprecipitation), 0018 (two hybrid), 0019 (coimmunoprecipitation), 0027 (cosedimentation), 0045 (experimental interaction detection), 0096 (pull down), 0416 (fluorescence microscopy), 0424 (protein kinase assay), 0435 (protease assay), 0515 (methyltransferase assay), 0889 (acetylation assay). For the NCBI database, we used acetylated by, acetylates, binds, cleaved by, cleaves, degraded by, degrades, dephosphorylates, methylated by, myristoylated by, phosphorylated by, phosphorylates, ubiquitinated by edges. We gathered HIV siRNA results from two sources covering four studies, and converted these hits to Entrez Gene IDs. Three studies were summarized in Table 4 of the supplementary document supplied by Bushman et al. at http://www.hostpathogen.org [24]. The fourth study was taken from Figure S2 in an article by Yeung et al. [181].

HCV-human protein interactions were taken from de Chassey et al. [40]. HCV

siRNA results were taken as genes that showed a decrease in infection, listed in supplementary material SD1 and SD2 columns B-E [102].

Influenza-human protein interactions and siRNA hits were taken from a study of host factors involved in influenza replication [91]. For the protein interactions, we ignored interactions where no virus protein was specified.

4.5.3 Peptide motif and SMART domain annotations

Protein sequences form three human interactomes, I2D, STRING, and HPRD, were scanned for peptide motifs and SMART domains. Protein sequences for the STRING and HPRD interactomes were taken from their respective databases. Protein sequences for the I2D network were taken from UniProt. Human proteins were scanned for SMART domains using batch access. Human proteins were annotated with the 136 peptide motifs described in the ELM Resource by downloading regular expressions for each motif from the resource, and matching them against all human protein sequences. The networks used in this study were composed of Entrez Gene IDs, and multiple proteins may correspond to one Entrez Gene ID. For peptide motif and SMART domain annotations for each Entrez Gene ID, we averaged the annotations of all the proteins for which it coded.

Chapter 5

Reflections and perspectives

In this dissertation, we presented three projects that introduced new observations about the nature of virus-host networks and generated testable hypotheses for further virus-host network discoveries. In the first project, we showed that host pathways targeted by HIV could be predicted using peptide motifs on HIV sequences. Our HIV-human interaction models have predicted new pathways and interactions that may be important for HIV infection. In our second project, we examined the docking between HIV proteins and human mitogen-activated protein kinases, which may be important for HIV replication, and proposed docking sites on HIV substrates that can be evaluated in the lab. In our third project, we contributed to the network biology field by addressing the observation that viruses target host network hub proteins, and asking if viruses had a preferential interaction with intermodular or intramodular hubs. By demonstrating a preference of intramodular hubs over intermodular hubs for HIV and influenza virus, we aided the systems study of biological networks by providing more evidence for the distinction between intermodular and intramodular hubs, which is currently under debate [1, 12, 13, 67, 162]. Furthermore, the virus intramodular hub preference promotes the study of which intramodular hub properties are important for viral infection. In this final chapter, we review the work presented in this dissertation and address how it can be used as a basis for future

investigations of virus-host networks.

5.1 Review of our work

Prediction of HIV virus-host protein interactions using virus and host sequence motifs

In Chapter 2 we presented a peptide motif based virus-host interaction prediction method, and tested its ability to accurately recover HIV-human interactions listed in the NCBI HIV-Human Protein Interaction Database [62, 130]. We motivated our work by outlining the importance of predicted virus-host interactions for guiding experimental studies of virus-host interactions [79, 96, 153]. The virus-host networks that emerged from virus-host interaction experiments have been used to annotate virus proteins of unknown function and compare different viral strategies for dealing with the host immune system [26, 116]. We showed that our predicted interactions had significant overlap with interactions in the NCBI database, and that virus targeted proteins from our predictions overlapped significantly with a set of host proteins that are important for HIV replication [24]. Our HIV-human interaction predictions were further validated in that the human proteins in the interactions occupied many of the same biological pathways as the human proteins shown to be targeted by HIV in the validated NCBI interactions. We showed that our predicted virus targeted proteins were also enriched in some pathways not known to interact with HIV, providing new potential directions for the study of virus-host networks. Our work for this chapter has been further summarized in a review by Chan et al. [28].

Our prediction work has generated hypotheses about new virus targeted host pathways and provided a list of host proteins whose interactions with virus proteins may be essential for HIV replication. The cell cycle, Jak-STAT, cytoskeletal regulation, and tight junction KEGG pathways were all significantly enriched in

our predicted interacting proteins for some HIV protein, but the corresponding enrichment was not significant for the validated virus-host interactions from NCBI. These pathways offer new hypotheses for cell processes that HIV might need to target. Combining the predicted virus-host interactions with the results form siRNA screens searching for host factors that are important for HIV replication also leads to new hypotheses. Predicted virus targeted proteins that are implicated in an siRNA screen can be tested to see if preventing the interaction has an effect on HIV replication. Furthermore, we have identified several protein binding sites on host and virus proteins that may be guiding HIV-human interactions that are essential for replication. Following up on the proposed virus-host interactions that might be important for replication and the suggested virus targeted pathways will help construct a more complete HIV-human interaction network that can facilitate future HIV studies.

A bioinformatics approach reveals possible MAPK docking motifs on HIV proteins

In Chapter 3 we continued our work with the hypothesis that virus proteins use host peptide motifs to interact with host proteins, and focused on the peptide motif that acts as a substrate docking site for mitogen-activated protein kinases (MAPKs) ERK1 and ERK2 [9]. Our work with MAPK and virus proteins was motivated by the importance of MAPK phosphorylation of HIV substrates MA and Vif in infection [22, 180]. We observed that HIV proteins MA, Rev, Tat, and Vif, while documented to be phosphorylated by ERK1 and ERK2, were missing the accepted MAPK docking motif. The HIV Nef protein had hits for the MAPK docking motif pattern, but further investigation of Nef's structure suggested that these sites were not functional. We revealed that modifications of the accepted MAPK docking motif pattern would yield peptide motif patterns that annotated all HIV proteins phosphorylated by ERK1 and ERK2. As an argument that our proposed docking motifs were functional, we showed that they were enriched on human MAPK ERK1

and ERK2 substrates, and we demonstrated that *in silico* docking of MAPK ERK1 and HIV MA via the proposed docking site produced a protein complex that aligned the active site of ERK1 with a possible phosphorylation site on MA. The locations of our proposed docking motifs on HIV proteins serve as testable sites that mediate MAPK and HIV protein interactions. If functional, our proposed docking sites will aid in the search for small-molecule drugs that prevent HIV protein phosphorylation by MAPKs ERK1 and ERK2.

The work in this chapter is important for future work with virus-host interactions. First, it serves as outline of computational steps for the modification and verification of host peptide motif for use on virus proteins. Second, this chapter motivates more questions concerning the presence of host peptide motifs on virus proteins. Are there other peptide motifs, like the MAPK docking site, that are variations of documented host peptide motifs? How is the virus utilization of variant host peptide motif patterns beneficial to the virus? Answering these questions will yield more insights into the nature of virus-host networks.

Modularity in protein interaction network hubs predicts viral host-pathogen interactions

In Chapter 4 we bridged the gap between studies of hubs in single organism networks and work done with virus-host networks to determine if host hub modularity, i.e., the presence of two hub types in networks, played a role in virus-host interactions. We reaffirmed the debated existence of intermodular and intramodular hubs in single organism networks using three human protein-protein interaction networks, and confirmed some of the properties that have been observed for intermodular and intramodular hubs, such as the preference of intramodular hubs to be parts of proteins complexes [59]. We showed that the despite being debated [1, 12, 13, 67, 162], the inter/intramodular hub distinction is important for network systems biology by demonstrating that HIV and influenza virus proteins have a significant interaction

preference for intramodular hubs. We ended this chapter by proposing that the virus intramodular hub preference is caused by a virus preference to interact with hubs that are part of host protein complexes.

The virus intramodular hub preference described in this work promotes questions about the features of host proteins that are targeted by viruses. Intramodular hubs evolve at slower rates [60], and are more structured than intermodular hubs [52, 152, 165]. Perhaps viruses are targeting these hub features in addition to the intramodular hub complex feature. More work should be done to investigate the importance of these intramodular hub features in guiding virus protein preference.

5.2 Future work

The work outlined here will help with future studies of virus-host networks. In addition to the experimental work suggested above, our work motivates other computational studies. Here we introduce four promising computational extensions to this work.

Predicting virus-host integrations using peptide motifs and additional information

One of the draw backs of predicting virus-host interactions using peptide motifs is the high number of false positive predictions. This problem can be alleviated by including additional information. Some of this information is provided by virus protein structures. HIV protein structures have already been used to predict virus-human interactions [46]. The peptide motif method can be supplemented by this work where virus protein structures are available. The peptide motif method can also easily fit into a prediction method that utilizes virus-host network motifs [172]. Network motifs are over-represented patterns of interaction involving two or more

proteins. Network motifs have been successfully used to predict yeast protein interactions [2]. It is likely that combining a similar approach using virus-host network motifs with the peptide motif based interaction prediction method will be able to predict virus-host interactions with fewer false positives.

Database of HIV mutations and their effects on virus-host interactions

Another fault of the peptide motif based virus-host interaction prediction method is the lack of evidence that an individual peptide sequence is responsible for a virus-host protein interaction. In the first aim of the dissertation, we showed that the conservation of peptide motifs on the HIV Nef protein sequences was not due to chance, and used this significant conservation to argue that conserved peptide motifs on HIV proteins were mediating virus-host interactions. In the second aim of the dissertation, we proposed that the statistical enrichment of a peptide motif on the interaction neighbors of a protein was evidence that the peptide motif was being used in interactions with some neighbor proteins.

The task of finding functional peptide motifs would be simplified if there was a database of mutations on virus proteins and their effects on virus-host interactions, identifying the functional motif hits would be much easier. The NCBI HIV-Human Interaction Database has some of the information needed to make such a database, but it is poorly organized. A database of HIV protein mutations can be built by taking all the source articles from the NCBI database, mining them for paragraphs mentioning mutations and virus-host interactions, and then having a large community of biologists annotation of these paragraphs with information describing the mutation and its effect on virus-host interactions. Using these virus mutations to find functional peptide motifs, and using functional peptide motifs with structure and network guides would make a better model of virus-host interactions.

Investigating the role of virus proteins as intermodular hubs in host protein interaction networks

Highly connected proteins in the human protein interaction network come in two types, intermodular hubs that modulate the activities of human complexes, and intramodular hubs that play important roles in these complexes [67, 162]. When combined with other studies of virus proteins, the work in this dissertation motivates the hypothesis that virus proteins act as intermodular hub proteins in virus-host networks. There is much evidence to support this claim. First, a study of influenza-human interactions revealed that influenza proteins had more interactions with host proteins than expected by chance, indicating that virus proteins might be hub proteins in the virus-host interaction network [146]. Second, in Chapter 4 we demonstrated that HIV and influenza proteins prefer to interact with intramodular hubs. We suggested that this hub preference was actually a preference to interact with host protein complexes, which is the same preference seen for intermodular hubs. Third, intermodular hubs are highly unstructured, or disordered, proteins compared to intramodular hubs [52, 152]. It has been observed that 25 RNA virus proteins with structures available in the Protein Data Bank [15] had large regions of protein disorder [165]. Specific cases of virus protein disorder involved the HIV Rev and matrix proteins [64, 170]. The basic protein segment of HIV Rev that binds HIV RNA is unstructured when alone in solution, but adopts an α-helix when bound to RNA [170]. The HIV matrix protein is highly disordered, and this might help the virus in evading the immune system [64]. Fourth, intermodular hubs have more peptide motifs than intramodular hubs [162]. In Chapter 2, we showed that not only do virus proteins have many conserved host peptide motifs, but these motifs are possibly functional because they can be used to predict interactions with virus proteins.

Testing the hypothesis that virus proteins act as intermodular hubs in the virus-host network is important because it has implications for antiviral therapies. Disordered proteins must be tightly regulated in host cells because their binding promiscuity makes then highly sensitive to changes in their concentration [66, 173]. There is evidence that concentration changes also affect HIV infection. In an HIV infected cell, HIV proteins Env, Nef, and Vpu regulate the presence of the HIV CD4 receptor at the cell's membrane [100]. Downregulating the HIV CD4 receptor at the cell's membrane is thought to decrease the chance of reinfection by more HIV virions, which would stress the host cell's pathways without resulting in the production of more virus particles [6, 71]. CD4 downregulation by HIV hints that the tight regulation of HIV protein expression is necessary for successful replication. More extensive investigation is needed to validate this hypothesis and develop methods to alter virus protein regulation for therapy purposes.

Examining the effects of host environment on influenza virus peptide motif usage

Influenza virus has the ability to infect a number of host organisms, including human, chicken, swine, and horse. Nucleotide sequences of the 2009 H1N1 pandemic influenza have been shown to have a substitution bias that depends on the host organism in which it resides [155]. Based on our work in Chapters 2 and 3, we propose a project to examine the possibility of an influenza virus peptide motif usage bias that correlates with host organism. Our results in Chapter 2 suggested that peptide motifs on virus proteins are important for guiding virus-host protein interactions. Case studies focusing on single peptide motifs on certain virus proteins have shown that some virus proteins use host peptide motifs to interact with host proteins to accomplish necessary steps in the viral life-cycle [82]. Regions of some influenza proteins have already been found to evolve differently in different hosts. The identities of four amino acids in an influenza polymerase component have been found to differ

consistently between mammalian and avian hosts, and it has been suggested that these differences affect interactions with host proteins [178]. In light of the importance of peptide motifs for virus-host interactions, we suggest conducting a study of differential peptide motif usage among influenza viruses infecting mammalian and avian hosts. Such a study will give insights into the selective pressures on influenza proteins, which can aid in drug design and the determination of which organism an influenza strain has originated.

Appendix A

Supplemental tables

HIV protein	Alignment sequence count
CA	824
ENV	411
GAG	824
IN	74
MA	824
NC	824
NEF	807
POL	74
PR	74
REV	417
RT	74
TAT	338
VIF	673
VPR	571
VPU	285

Table A.1: This table shows the number of protein sequences in each multiple alignment for all HIV proteins.

ELM	Binding PROSITE or Proteins
CLV_NDR_NDR_1	NP_002516.1
CLV_PCSK_FUR_1	NP_002560.1
CLV_PCSK_PC1ET2_1	NP_002585.2; NP_000430.3
CLV_PCSK_PC7_1	NP_004707.2
CLV_PCSK_SKI1_1	NP_056453.1
LIG_14-3-3_3	14-3-3 proteins signature 1
LIG_14-3-3_3	14-3-3 proteins signature 2
LIG_APCC_Dbox_1	BAA88957.1; NP_001246.2
LIG_BRCT_MDC1_1	BRCT domain profile
LIG_CYCLIN_1	Cyclins signature
LIG_Clathr_ClatBox_1	NP_004850.1
LIG_EH1_1	Trp-Asp (WD) repeats profile
LIG_EH1_1	Trp-Asp (WD) repeats circular profile
LIG_EH1_1	Trp-Asp (WD) repeats signature
LIG_EVH1_I	WH1 domain profile
LIG_FHA_1	Forkhead-associated (FHA) domain profile
LIG_FHA_2	Forkhead-associated (FHA) domain profile
LIG_MAPK_1	MAP kinase signature
LIG_NRBOX	Nuclear hormone receptors DNA-binding domain profile
LIG_NRBOX	Nuclear hormones receptors DNA-binding region signature
LIG_PDZ_3	PDZ domain profile
LIG_PP1	Serine/threonine specific protein phosphatases signature
LIG_PP2B_1	AAH28049.1;NP_671709.1;NP_000936.1;NP_000935.1;NP_005596.2
LIG_SH2_GRB2	Src homology 2 (SH2) domain profile
LIG_SH2_PTP2	Src homology 2 (SH2) domain profile
LIG_SH2_SRC	Src homology 3 (SH3) domain profile
LIG_SH2_STAT3	Src homology 2 (SH2) domain profile
LIG_SH2_STAT5	Src homology 2 (SH2) domain profile
LIG_SH3_1	Src homology 3 (SH3) domain profile
LIG_SH3_2	Src homology 3 (SH3) domain profile
LIG_SH3_3	Src homology 3 (SH3) domain profile
LIG_TRAF2_1	MATH/TRAF domain profile
LIG_TRFH_1	NP_059523.1; NP_005643.1
LIG_ULM_U2AF65_1	Eukaryotic RNA Recognition Motif (RRM) profile
LIG_USP7_1	NP_003461.1

Table A.2: We associated each peptide motif (ELM) with an interacting domain or protein set (CD). This table has been truncated from the full version, which is available at http://www.biomedcentral.com/content/supplementary/1755-8794-2-27-s2.xls. The full version shows the fraction of human proteins used in the study that are annotated with an ELM or its interacting CD. It also shows the faction of protein interactions that satisfy the ELM-CD relation.

VP	H1	DHHE	Match	Precision	Recall	Pvalue
ENV	527	177	25	4.7438330170778	14.1242937853107	6.57E-003
GAG	404	53	7	1.73267326732673	13.2075471698113	2.51E-002
IN	398	45	3	0.753768844221106	6.66666666666667	3.54E-001
MA	196	29	4	2.04081632653061	13.7931034482759	2.29E-003
NEF	399	55	12	3.00751879699248	21.8181818181818	4.92E-005
POL	482	114	18	3.7344398340249	15.7894736842105	1.81E-003
PR	203	49	0	0	0	1.00
REV	354	27	4	1.12994350282486	14.8148148148148	1.97E-002
RT	388	17	13	3.35051546391753	76.4705882352941	1.13E-014
TAT	188	213	24	12.7659574468085	11.2676056338028	8.93E-009
VIF	450	35	1	0.222222222222222	2.85714285714286	7.54E-001
VPR	178	24	1	0.561797752808989	4.16666666666667	1.60E-001

Table A.3: Here we compare predicted (H1) and experimentally verified (DHHE) direct virus-host interactions for all HIV proteins, giving the overlap (Match) between the two protein sets. P-values are calculated as the probability of matching Match genes or more when comparing DHHE and H1 drawn from the 5954 proteins in the study (see Methods).

Pathway type	VP	HHP	HHE	Match	Pvalue
All Pathways	TAT	621	509	150	2.95639225843e-37
HHP Enriched Pathways	TAT	410	509	112	1.57137483607e-32
All Pathways	PR	560	58	19	1.27382356047e-07
HHP Enriched Pathways	PR	310	58	14	1.55046126572e-07
All Pathways	MA	589	47	16	1.06498484182e-06
HHP Enriched Pathways	MA	323	47	15	1.26127368289e-09
All Pathways	VPR	444	119	36	1.30088275718e-14
HHP Enriched Pathways	VPR	202	119	26	1.03299312653e-15
All Pathways	NEF	866	155	69	1.4369577924e-20
HHP Enriched Pathways	NEF	519	155	54	4.26449335629e-21
All Pathways	CA	853	7	1	0.264613621249
HHP Enriched Pathways	CA	529	7	0	1.0
All Pathways	NC	6	7	0	1.0
HHP Enriched Pathways	NC	5	7	0	1.0
All Pathways	REV	797	40	11	0.00470778327471
HHP Enriched Pathways	REV	479	40	7	0.0127983863936
All Pathways	POL	991	122	45	1.37533161737e-08
HHP Enriched Pathways	POL	618	122	38	4.10000804347e-11
All Pathways	VPU	81	7	1	0.00367272286524
HHP Enriched Pathways	VPU	34	7	1	0.000652943299923
All Pathways	ENV	1013	409	170	5.48212311982e-35
HHP Enriched Pathways	ENV	584	409	127	7.703084244e-37
All Pathways	IN	871	46	5	0.683215864589
HHP Enriched Pathways	IN	520	46	2	0.778834368003
All Pathways	GAG	938	103	36	3.6204620022e-07
HHP Enriched Pathways	GAG	548	103	25	1.22862968376e-06
All Pathways	VIF	888	35	5	0.425490399203
HHP Enriched Pathways	VIF	520	35	1	0.823304149319
All Pathways	RT	911	23	19	4.61361299622e-14
HHP Enriched Pathways	RT	518	23	19	6.06244810602e-19

Table A.4: Predicted (HHP) and experimentally validated (HHE) HIV-human virus-host interactions are compared for all HIV proteins when human proteins in predicted virus-host interations are restricted to genes in all KEGG pathways, and KEGG pathways enriched (p-value < 0.01, see Methods) with our predictions. P-values are calculated as the probability of matching Match genes or more when comparing HHE and HHP drawn from the 5954 proteins in the study (see Methods).

Bibliography

[1] S. Agarwal, C.M. Deane, M.A. Porter, and N.S. Jones. Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Computational Biology*, 6(6):e1000817, 2010.

[2] I. Albert and R. Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346, 2004.

[3] R. Albert, H. Jeong, and A.L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[5] M.R. Arkin and J.A. Wells. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery*, 3(4):301–317, 2004.

[6] S.T. Arold and A.S. Baur. Dynamic Nef and Nef dynamics: how structure could explain the complex activities of this small HIV protein. *Trends in Biochemical Sciences*, 26(6):356–363, 2001.

[7] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[8] A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[9] A. J. Bardwell, E. Frankson, and L. Bardwell. Selectivity of docking sites in MAPK kinases. *Journal of Biological Chemistry*, 284(19):13165, 2009.

[10] L. Bardwell. Mechanisms of MAPK signalling specificity. *Biochemical Society Transactions*, 34:837–841, 2006.

[11] P. Barraud, J. C. Paillart, R. Marquet, and C. Tisne. Advances in the structural understanding of Vif proteins. *Current HIV Research*, 6(2):91–9, 2008.

[12] N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, L. D. Hurst, and M. Tyers. Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biology*, 4(10):e317, 2006.

[13] N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, L. D. Hurst, and M. Tyers. Still stratus not altocumulus: further evidence against the date/party hub distinction. *PLoS Biology*, 5(6):e154, 2007.

[14] N.N. Batada, L.D. Hurst, and M. Tyers. Evolutionary and physiological importance of hub proteins. *PLoS Comput Biol*, 2(7):e88, 2006.

[15] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, and S. Jain. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

[16] N. Bertin, N. Simonis, D. Dupuy, M. E. Cusick, J. D. J. Han, H. B. Fraser, F. P. Roth, and M. Vidal. Confirmation of organized modularity in the yeast interactome. *PLoS Biology*, 5(6), 2007.

[17] J. Blenis. Signal transduction via the MAP kinases: proceed at your own RSK. *PNAS*, 90(13):5889, 1993.

[18] C. Brander and B.D. Walker. Modulation of host immune responses by clinically relevant human DNA and RNA viruses. *Current Opinion in Microbiology*, 3(4):379–386, 2000.

[19] A. L. Brass, D. M. Dykxhoorn, Y. Benita, N. Yan, A. Engelman, R. J. Xavier, J. Lieberman, and S. J. Elledge. Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, 319(5865):921, 2008.

[20] J. N. Brown, J. J. Kohler, C. R. Coberley, J. W. Sleasman, and M. M. Goodenow. HIV-1 activates macrophages independent of toll-like receptors. *PLoS ONE*, 3(12):e3664, 2008.

[21] K.R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076, 2005.

[22] A. G. Bukrinskaya, A. Ghorpade, N. K. Heinzinger, T. E. Smithgall, R. E. Lewis, and M. Stevenson. Phosphorylation-dependent human immunodeficiency virus type 1 infection and nuclear targeting of viral DNA, 1996.

[23] K. Burkhard, S. Smith, R. Deshmukh, A. D. MacKerell Jr, and P. Shapiro. Development of extracellular signal-regulated kinase inhibitors. *Current Topics in Medicinal Chemistry*, 9(8):678–689, 2009.

[24] F.D. Bushman, N. Malani, J. Fernandes, I. D'Orso, G. Cagney, T.L. Diamond, H. Zhou, D.J. Hazuda, A.S. Espeseth, R. König, et al. Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathogens*, 5(5), 2009.

[25] R. Byland, P. J. Vance, J. A. Hoxie, and M. Marsh. A conserved dileucine motif mediates clathrin and AP-2-dependent endocytosis of the HIV-1 envelope protein. *Molecular Biology of the Cell*, 18(2):414, 2007.

[26] M. A. Calderwood, K. Venkatesan, L. Xing, M. R. Chase, A. Vazquez, A. M. Holthaus, A. E. Ewence, N. Li, T. Hirozane-Kishikawa, and D. E. Hill. Epstein-Barr virus and virus human protein interaction maps. *PNAS*, 104(18):7606, 2007.

[27] A. Ceol, A. Chatr-aryamontri, E. Santonico, R. Sacco, L. Castagnoli, and G. Cesareni. DOMINO: a database of domain-peptide interactions. *Nucleic Acids Research*, 2006.

[28] Eric Y Chan, Marcus J Korth, and Michael G Katze. Decoding the multi-faceted HIV-1 virus-host interactome. *Journal of biology*, 8(9):84, January 2009.

[29] A. Chatr-aryamontri, A. Ceol, D. Peluso, A. Nardozza, S. Panni, F. Sacco, M. Tinti, A. Smolyar, L. Castagnoli, M. Vidal, et al. VirusMINT: a viral protein interaction database. *Nucleic Acids Research*, 2008.

[30] R. Chen, L. Li, and Z. Weng. ZDOCK: an initial-stage protein-docking algorithm. *Proteins Structure Function and Genetics*, 52(1):80–87, 2003.

[31] S. Clancy. Genetics of the influenza virus. *Nature Education*, 1(1), 2008.

[32] S. H. Coleman, R. Madrid, N. Van Damme, R. S. Mitchell, J. Bouchet, C. Servant, S. Pillai, S. Benichou, and J. C. Guatelli. Modulation of cellular protein trafficking by human immunodeficiency virus type 1 Nef: Role of the acidic residue in the ExxxLL motif. *Journal of Virology*, 80(4):1837–1849, 2006.

[33] S.R. Collins, P. Kemmeren, X.C. Zhao, J.F. Greenblatt, F. Spencer, F.C.P. Holstege, J.S. Weissman, and N.J. Krogan. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, 6(3):439, 2007.

[34] The UniProt Consortium. The universal protein resource (UniProt) in 2010. *Nucl. Acids Res.*, 38(suppl_1):D142–148, January 2010.

[35] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E.D. Spear, C.S. Sevier, H. Ding, J.L.Y. Koh, K. Toufighi, S. Mostafavi, et al. The genetic landscape of a cell. *Science*, 327(5964):425, 2010.

[36] W. Dampier and A. Tozeren. Signaling perturbations induced by invading h. pylori proteins in the host epithelial cells: A mathematical modeling approach. *Journal of Theoretical Biology*, 248(1):130–144, 2007.

[37] William Dampier, Perry Evans, Lyle Ungar, and Aydin Tozeren. Host sequence motifs shared by HIV predict response to antiretroviral therapy. *BMC Medical Genomics*, 2(1):47, 2009.

[38] F. P. Davis, D. T. Barkan, N. Eswar, J. H. McKerrow, and A. Sali. Host pathogen protein interactions predicted by comparative modeling. *Protein Science*, 16(12):2585, 2007.

[39] E. de Castro, C. J. A. Sigrist, A. Gattiker, V. Bulliard, P. S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch, and N. Hulo. ScanProsite: detection of PROSITE signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Research*, 34(Web Server issue):W362, 2006.

[40] B. De Chassey, V. Navratil, L. Tafforeau, M. S. Hiet, A. Aublin-Gex, S. Agauguè, G. Meiffren, F. Pradezynski, B. F. Faria, and T. Chantier. Hepatitis c virus infection protein network. *Molecular Systems Biology*, 4(1), 2008.

[41] S. G. Deeks. Treatment of antiretroviral-drug-resistant HIV-1 infection. *The Lancet*, 362(9400):2002–2011, 2003.

[42] A. Degterev, A. Lugovskoy, M. Cardone, B. Mulley, G. Wagner, T. Mitchison,

and J. Yuan. Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-xL. *Nature Cell Biology*, 3(2):173–182, 2001.

[43] G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(3):2003–4, 2003.

[44] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N. P. Brown, G. Trave, and T. J. Gibson. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci*, 13:65806603, 2008.

[45] H. Dinkel and H. Sticht. A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics*, 23(24):3297, 2007.

[46] J. Doolittle and S. Gomez. Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. *Virology Journal*, 7(1):82, 2010.

[47] T. Driscoll, M.D. Dyer, TM Murali, and B.W. Sobral. PIG–the pathogen interaction gateway. *Nucleic acids research*, 37(Database issue):D647, 2009.

[48] J. Dutkowski and J. Tiuryn. Identification of functional modules from conserved ancestral protein protein interactions. *Bioinformatics*, 23(13):i149, 2007.

[49] M. D. Dyer, T. M. Murali, and B. W. Sobral. Computational prediction of host-pathogen protein protein interactions. *Bioinformatics*, 23(13):i159, 2007.

[50] M. D. Dyer, T. M. Murali, and B. W. Sobral. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathogens*, 4(2):e32, 2008.

[51] R. J. Edwards, N. E. Davey, and D. C. Shields. SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, 2(10):e967, 2007.

[52] D. Ekman, S. Light, A. Bjorklund, and A. Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biology*, 7(6):R45, 2006.

[53] A. Ertel and A. Tozeren. Switch-like genes populate cell communication pathways and are enriched for extracellular proteins. *BMC Genomics*, 9(1):3, 2008.

[54] P. Evans, W. Dampier, L. Ungar, and A. Tozeren. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Medical Genomics*, 2(1):27, 2009.

[55] N. Evrard-Todeschi, J. Gharbi-Benarous, G. Bertho, G. Coadou, S. Megy, R. Benarous, and J. P. Girault. NMR studies for identifying phosphopeptide ligands of the HIV-1 protein Vpu binding to the F-box protein beta-TrCP. *Peptides*, 27(1):194, 2006.

[56] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Research*, 36:D281–288, 2008.

[57] A. Fox, D. Taylor, and D. K. Slonim. High throughput interaction data reveals degree conservation of hub proteins. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 391, 2009.

[58] A.D. Frankel and J.A.T. Young. HIV-1: fifteen proteins and an RNA. *Annual Reviews Biochemistry*, 2003.

[59] H.B. Fraser. Modularity and evolutionary constraint on proteins. *Nature genetics*, 37(4):351–352, 2005.

[60] H.B. Fraser. Modularity and evolutionary constraint on proteins. *Nature Genetics*, 37(4):351–352, 2005.

[61] H.B. Fraser, A.E. Hirsh, L.M. Steinmetz, C. Scharfe, and M.W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750, 2002.

[62] W. Fu, B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt, and R. G. Ptak. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Research*, 37(Database issue):D417, 2009.

[63] P. Georgel, C. Schuster, M.B. Zeisel, F. Stoll-Keller, T. Berg, S. Bahram, and T.F. Baumert. Virus-host interactions in hepatitis C virus infection: implications for molecular pathogenesis and antiviral strategies. *Trends in Molecular Medicine*, 2010.

[64] G. Goh, A.K. Dunker, and V. Uversky. Protein intrinsic disorder toolbox for comparative analysis of viral proteins. *BMC Genomics*, 9(Suppl 2):S4, 2008.

[65] C.M. Gould, F. Diella, A. Via, P. Puntervoll, C. Gemund, S. Chabanis-Davidson, S. Michael, A. Sayadi, J.C. Bryne, C. Chica, et al. ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Research*, 38(Database issue):D167, 2010.

[66] J. Gsponer, M.E. Futschik, S.A. Teichmann, and M.M. Babu. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science*, 322(5906):1365, 2008.

[67] J.D.J. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J.M. Walhout, M.E. Cusick, F.P. Roth, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, 2004.

[68] J.D.J. Han, D. Dupuy, N. Bertin, M.E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844, 2005.

[69] C. N. Hancock, A. T. Macias, A. D. Mackerell Jr, and P. Shapiro. Mitogen activated protein (MAP) kinases: development of ATP and non-ATP dependent inhibitors. *Medicinal chemistry (Sh riqah (United Arab Emirates))*, 2(2):213, 2006.

[70] L. Hao, A. Sakurai, T. Watanabe, E. Sorensen, C.A. Nidom, M.A. Newton, P. Ahlquist, and Y. Kawaoka. Drosophila RNAi screen identifies host genes important for influenza virus replication. *Nature*, 454(7206):890–893, 2008.

[71] M. Harris. HIV: a new role for Nef in the spread of HIV. *Current Biology*, 9(12):R459–R461, 1999.

[72] G.T. Hart, A.K. Ramani, and E.M. Marcotte. How complete are current yeast and human protein-interaction networks. *Genome Biol*, 7(11):120, 2006.

[73] X. He and J. Zhang. Why do hubs tend to be essential in protein networks. *PLoS Genetics*, 2(6):e88, 2006.

[74] J. Hemelaar, E. Gouws, P. D. Ghys, and S. Osmanov. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *Aids*, 20(16):W13, 2006.

[75] E. C. Holmes. When HIV spread afar. *PNAS*, 104(47):18351, 2007.

[76] H. Huang and J.S. Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3):372, 2009.

[77] E. Hubbell, W.M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585, 2002.

[78] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B. A. Cuche, E. de Castro, C. Lachaize, P. S. Langendijk-Genevaux, and C. J. A. Sigrist. The 20 years of PROSITE. *Nucleic Acids Research*, 36(Database issue):D245, 2008.

[79] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449, 2003.

[80] H. Jeong, S.P. Mason, A.L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[81] P.F. Jonsson and P.A. Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291, 2006.

[82] K. Kadaveru, J. Vyas, and M. R. Schiller. Viral infection and human disease-insights from minimotifs. *Frontiers in Bioscience*, 13:6455–71, 2008.

[83] R. Kafri, O. Dahan, J. Levy, and Y. Pilpel. Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *PNAS*, 105(4):1243, 2008.

[84] B. Kahali, S. Ahmad, and T.C. Ghosh. Exploring the evolutionary rate differences of party hub and date hub proteins in Saccharomyces cerevisiae protein-protein interaction network. *Gene*, 429(1-2):18–22, 2009.

[85] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, and T. Tokimatsu. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):D480, 2008.

[86] U. Karaoz, TM Murali, S. Letovsky, Y. Zheng, C. Ding, C.R. Cantor, and S. Kasif. Whole-genome annotation by using evidence integration in functional-linkage networks. *PNAS*, 101(9):2888, 2004.

[87] B.P. Kelley, R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell, and

T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS*, 100(20):11394, 2003.

[88] P.M. Kim, A. Sboner, Y. Xia, and M. Gerstein. The role of disorder in interaction networks: a structural analysis. *Molecular Systems Biology*, 4(1), 2008.

[89] T. Kinoshita, I. Yoshida, S. Nakae, K. Okita, M. Gouda, M. Matsubara, K. Yokota, H. Ishiguro, and T. Tada. Crystal structure of human monophosphorylated ERK1 at Tyr204. *Biochemical and Biophysical Research Communications*, 377(4):1123–1127, 2008.

[90] W. Kolch. Coordinating ERK/MAPK signalling through scaffolds and inhibitors. *Nature Reviews Molecular Cell Biology*, 6(11):827–837, 2005.

[91] R. König, S. Stertz, Y. Zhou, A. Inoue, H.H. Hoffmann, S. Bhattacharyya, J.G. Alamares, D.M. Tscherne, M.B. Ortigoza, Y. Liang, et al. Human host factors required for influenza virus replication. *Nature*, 463(7282):813–817, 2009.

[92] R. König, Y. Zhou, D. Elleder, T. L. Diamond, G. M. C. Bonamy, J. T. Irelan, C. Chiang, B. P. Tu, P. D. De Jesus, and C. E. Lilley. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1):49–60, 2008.

[93] N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[94] T. Kuntzen, J. Timm, A. Berical, L. L. Lewis-Ximenez, A. Jones, B. Nolan, J. S. zur Wiesch, B. Li, A. Schneidewind, and A. Y. Kim. Viral sequence evolution in acute hepatitis c virus infection. *Journal of Virology*, 81(21):11658, 2007.

[95] J.N. Lavoie, G. L'Allemain, A. Brunet, R. Muller, and J. Pouyssegur. Cyclin D1 expression is regulated positively by the p42/p44MAPK and negatively by the p38/HOGMAPK pathway. *Journal of Biological Chemistry*, 271(34):20608, 1996.

[96] I. Lee, S.V. Date, A.T. Adai, and E.M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555, 2004.

[97] T. Y. Lee, H. D. Huang, J. H. Hung, H. Y. Huang, Y. S. Yang, and T. H. Wang. dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Research*, 34(Database Issue):D622, 2006.

[98] T. Lengauer and T. Sing. Bioinformatics-assisted anti-HIV therapy. *Nature Reviews Microbiology*, 4(10):790–797, 2006.

[99] I. Letunic, T. Doerks, and P. Bork. SMART 6: recent updates and new developments. *Nucleic Acids Research*, 2008.

[100] K. Levesque, A. Finzi, J. Binette, and EA Cohen. Role of CD4 receptor down-regulation during HIV-1 infection. *Current HIV Research*, 2(1):51–59, 2004.

[101] D. Li, J. Li, S. Ouyang, J. Wang, S. Wu, P. Wan, Y. Zhu, X. Xu, and F. He. Protein interaction networks of Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster: large-scale organization and robustness. *Proteomics*, 6(2):456–461, 2006.

[102] Q. Li, A.L. Brass, A. Ng, Z. Hu, R.J. Xavier, T.J. Liang, and S.J. Elledge. A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *PNAS*, 106(38):16410, 2009.

[103] H. Y. Lin, S. L. Zhang, B. L. West, H. Y. Tang, T. Passaretti, F. B. Davis, and

P. J. Davis. Identification of the putative MAP kinase docking site in the thyroid hormone receptor-beta 1 DNA-binding domain: Functional consequences of mutations at the docking site. *Biochemistry*, 42(24):7571–7579, 2003.

[104] S. LIU, J.P. SUN, BO ZHOU, and Z.Y. ZHANG. Structural basis of docking interactions between ERK2 and MAP kinase phosphatase 3. *PNAS*, 103(14):5326–5331, 2006.

[105] Y. Liu and A. Tozeren. Modular composition predicts kinase/substrate interactions. *BMC bioinformatics*, 11(1):349, 2010.

[106] X. Lu, V.V. Jain, P.W. Finn, and D.L. Perkins. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Molecular Systems Biology*, 3(1), 2007.

[107] W. Lv, Z. Liu, H. Jin, X. Yu, and L. Zhang. Three-dimensional structure of HIV-1 Vif constructed by comparative modeling and the function characterization analyzed by molecular dynamics simulation. *Organic & Biomolecular Chemistry*, 5(4):617–626, 2007.

[108] L.M. Mansky. Retrovirus mutation rates and their role in genetic variation. *Journal of general Virology*, 79(6):1337, 1998.

[109] C.J. Marshall. MAP kinase kinase kinase, MAP kinase kinase and MAP kinase. *Current opinion in genetics & development*, 4(1):82–89, 1994.

[110] A. Mehle, E. R. Thomas, K. S. Rajendran, and D. Gabuzda. A zinc-binding region in Vif binds Cul5 and determines cullin selection. *Journal of Biological Chemistry*, 281(25):17259, 2006.

[111] J. Mendez-Rios and P. Uetz. Global approaches to study protein–protein interactions among viruses and hosts. *Future Microbiology*, 5(2):289–301, 2010.

[112] N. Misawa, AKM Kafi, M. Hattori, K. Miura, K. Masuda, and T. Ozawa. Rapid and high-sensitivity cell-based assays of protein-protein interactions using split click beetle luciferase complementation: An approach to the study of G-protein-coupled receptors. *Analytical Chemistry*, 82(6):2552–2560, 2010.

[113] D. Moradpour, F. Penin, and C.M. Rice. Replication of hepatitis C virus. *Nature Reviews Microbiology*, 5(6):453–463, 2007.

[114] J. R. Morgan, K. Prasad, S. Jin, G. J. Augustine, and E. M. Lafer. Eps15 homology domain-NPF motif interactions regulate clathrin coat assembly during synaptic vesicle recycling. *Journal of Biological Chemistry*, 278(35):33583–33592, 2003.

[115] D.J. Munroe and T.J.R. Harris. Third-generation sequencing fireworks at Marco Island. *Nature Biotechnology*, 28(5):426–428, 2010.

[116] V. Navratil, B. de Chassey, L. Meyniel, F. Pradezynski, P. Andre, C. Rabourdin-Combe, and V. Lotteau. System-level comparison of protein-protein interactions between viruses and the human type I interferon system network. *Journal of Proteome Research*, pages 1–20, 2010.

[117] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS BIOLOGY*, 3(12):2090, 2005.

[118] V. Neduva and R. B. Russell. Linear motifs: Evolutionary interaction switches. *FEBS Letters*, 579(15):3342–3345, 2005.

[119] V. Neduva and R. B. Russell. DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Research*, 34(Web Server issue):W350, 2006.

[120] V. Neduva and R. B. Russell. Peptides mediating interaction networks: new leads at last. *Current Opinion in Biotechnology*, 17(5):465–471, 2006.

[121] V. Neduva and R.B. Russell. Peptides mediating interaction networks: new leads at last. *Current opinion in biotechnology*, 17(5):465–471, 2006.

[122] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Research*, 2007.

[123] C. Pan, J. V. Olsen, H. Daub, and M. Mann. Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics. *Molecular & Cellular Proteomics*, page M900285, 2009.

[124] G. Z. Panos, E. Xirouchakis, V. Tzias, G. Charatsis, I. A. Bliziotis, V. Doulgeroglou, N. Margetis, and M. E. Falagas. Helicobacter pylori infection in symptomatic HIV-seropositive and-seronegative patients: A case-control study. *AIDS Research and Human Retroviruses*, 23(5):709–712, 2007.

[125] V. R. Panz and B. I. Joffe. Impact of HIV infection and AIDS on prevalence of type 2 diabetes in South Africa in 2010, 1999.

[126] L. Parthasarathi, F. Casey, A. Stein, P. Aloy, and D.C. Shields. Approved drug mimics of short peptide ligands from protein interaction motifs. *J. Chem. Inf. Model*, 48(10):1943–1948, 2008.

[127] P. Patel, D. L. Hanson, P. S. Sullivan, R. M. Novak, A. C. Moorman, T. C. Tong, S. D. Holmberg, and J. T. Brooks. Incidence of types of cancer among HIV-infected persons compared with the general population in the united states, 1992-2003. *Annals of Internal Medicine*, 148(10):728, 2008.

[128] E. Petsalaki and R.B. Russell. Peptide-mediated interactions in biological

systems: new discoveries and applications. *Current Opinion in Biotechnology*, 19(4):344–350, 2008.

[129] T. S. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, and A. Venugopal. Human protein reference database–2009 update. *Nucleic Acids Research*, 2008.

[130] R. G. Ptak, W. Fu, B. E. Sanders-Beer, J. E. Dickerson, J. W. Pinney, D. L. Robertson, M. N. Rozanov, K. S. Katz, D. R. Maglott, and K. D. Pruitt. Cataloguing the HIV type 1 human protein interaction network. *AIDS Research and Human Retroviruses*, 24(12):1497–1502, 2008.

[131] P. Puntervoll, R. Linding, C. Gemnd, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. A. Martin, G. Ausiello, B. Brannetti, and A. Costantini. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research*, 31(13):3625–3630, 2003.

[132] J. Rachlin, D.D. Cohen, C. Cantor, and S. Kasif. Biological context networks: a mosaic view of the interactome. *Molecular Systems Biology*, 2(1), 2006.

[133] B. S. Ramakrishna. Prevalence of intestinal pathogens in HIV patients with diarrhea: Implications for treatment. *Indian Journal of Pediatrics*, 66(1):85–91, 1999.

[134] A. Rambaut, D. Posada, K.A. Crandall, and E.C. Holmes. The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 5(1):52–61, 2004.

[135] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, 35(Web Server issue):W193, 2007.

[136] I. Remy, A. Galarneau, and S.W. Michnick. Detection and visualization of

protein interactions with protein fragment complementation assays. *Methods in Molecular Biology*, 185:447–460, 2002.

[137] A. Remnyi, M. C. Good, and W. A. Lim. Docking interactions in protein kinase and phosphatase networks. *Current Opinion in Structural Biology*, 16(6):676–685, 2006.

[138] D.J. Robbins, E. Zhen, M. Cheng, S. Xu, D. Ebert, and M.H. Cobb. MAP kinases ERK1 and ERK2: pleiotropic enzymes in a ubiquitous signaling network. *Advances in cancer research*, 63:93–116, 1994.

[139] M.H.A. Roehrl, S. Kang, J. Aramburu, G. Wagner, A. Rao, and P.G. Hogan. Selective inhibition of calcineurin-NFAT signaling by blocking protein–protein interaction with small organic molecules. *PNAS*, 101(20):7554, 2004.

[140] J. F. Roeth and K. L. Collins. Human immunodeficiency virus type 1 Nef: Adapting to intracellular trafficking pathways. *Microbiology and Molecular Biology Reviews*, 70(2):548, 2006.

[141] K. Roovers and R.K. Assoian. Integrating the MAP kinase signal into the G1 phase cell cycle machinery. *Bioessays*, 22(9):818–826, 2000.

[142] T. Ruths, D. Ruths, and L. Nakhleh. GS2: an efficiently computable measure of GO-based similarity of gene sets. *Bioinformatics*, 25(9):1178, 2009.

[143] A. Sacan, I. H. Toroslu, and H. Ferhatosmanoglu. Integrated search and alignment of protein structures. *Bioinformatics*, 24(24):2872, 2008.

[144] J. A. Schrager, V. Der Minassian, and J. W. Marsh. HIV Nef increases T cell ERK MAP kinase activity. *Journal of Biological Chemistry*, 277(8):6137, 2002.

[145] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(Suppl 1):i264, 2003.

[146] S.D. Shapira, I. Gat-Viks, B.O.V. Shum, A. Dricot, M.M. de Grace, L. Wu, P.B. Gupta, T. Hao, S.J. Silver, et al. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*, 139(7):1255–1267, 2009.

[147] R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974, 2005.

[148] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3(1), 2007.

[149] H. Shelton and M. Harris. Hepatitis C virus NS5 A protein binds the SH3 domain of the Fyn tyrosine kinase with high affinity: mutagenic analysis of residues within the SH 3 domain that contribute to the interaction. *Virology Journal*, 5(1):24, 2008.

[150] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

[151] C.W. Shepard, L. Finelli, M.J. Alter, et al. Global epidemiology of hepatitis C virus infection. *Lancet Infect Dis*, 5(9):558–567, 2005.

[152] G.P. Singh, M. Ganapathi, and D. Dash. Role of intrinsic disorder in transient interactions of hub proteins. *Proteins: Structure, Function, and Bioinformatics*, 66(4):761–765, 2007.

[153] L. Skrabanek, H.K. Saini, G.D. Bader, and A.J. Enright. Computational prediction of protein–protein interactions. *Molecular biotechnology*, 38(1):1–17, 2008.

[154] A. Sodhi, S. Montaner, and J. S. Gutkind. Viral hijacking of G-protein-coupled-receptor signalling networks. *Nature Reviews Molecular Cell Biology*, 5(12):998–1012, 2004.

[155] A. Solovyov, B. Greenbaum, G. Palacios, W.I. Lipkin, and R. Rabadan. Host Dependent Evolutionary Patterns and the Origin of 2009 H1N1 Pandemic Influenza. *PLoS Currents. Influenza*, 2010.

[156] M.P.H. Stumpf, C. Wiuf, and R.M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *PNAS*, 102(12):4221, 2005.

[157] P.H. Sugden and A. Clerk. Regulation of the ERK subgroup of MAP kinase cascades through G protein-coupled receptors. *Cellular signalling*, 9(5):337–351, 1997.

[158] H. Suzuki, Y. Fukunishi, I. Kagawa, R. Saito, H. Oda, T. Endo, S. Kondo, H. Bono, Y. Okazaki, and Y. Hayashizaki. Protein–protein interaction panel using mouse full-length cDNAs. *Genome Research*, 11(10):1758, 2001.

[159] S. H. Tan, W. Hugo, W. K. Sung, and S. K. Ng. A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, 7:502, 2006.

[160] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman. Prediction of interactions between HIV-1 and human proteins by information integration. In *Pacific Symposium on Biocomputing*, page 516, 2009. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.

[161] B. S. Taylor, M. E. Sobieszczyk, F. E. McCutchan, and S. M. Hammer. The challenge of HIV-1 subtype diversity. *New England Journal of Medicine*, 358(15):1590, 2008.

[162] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J.L. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27:199–204, 2009.

[163] K.M. Teshima and H. Innan. Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*, 178(3):1385, 2008.

[164] L.A. Thompson and J.A. Ellman. Synthesis and applications of small molecule libraries. *Chemical Reviews*, 96(1):555–600, 1996.

[165] N. Tokuriki, C.J. Oldfield, V.N. Uversky, I.N. Berezovsky, and D.S. Tawfik. Do viral proteins possess unique biophysical features? *Trends in Biochemical Sciences*, 34(2):53–59, 2009.

[166] R. Tonikian, Y. Zhang, S. L. Sazinsky, B. Currell, J. H. Yeh, B. Reva, H. A. Held, B. A. Appleton, M. Evangelista, and Y. Wu. A specificity map for the PDZ domain family. *PLoS Biology*, 6(9):e239, 2008.

[167] E. Toschi, I. Bacigalupo, R. Strippoli, C. Chiozzini, A. Cereseto, M. Falchi, F. Nappi, C. Sgadari, G. Barillari, and F. Mainiero. HIV-1 Tat regulates endothelial cell cycle progression via activation of the Ras/ERK MAPK signaling pathway. *Molecular Biology of the Cell*, 17(4):1985, 2006.

[168] J. N. Tournier and A. Quesnel-Hellmann. Host-pathogen interactions: A biological rendez-vous of the infectious nonself and danger models. *PLoS Pathogens*, 2(5):e44, 2006.

[169] R. Truant and B. R. Cullen. The arginine-rich domains present in human immunodeficiency virus type 1 Tat and Rev function as direct importin beta-dependent nuclear localization signals. *Molecular and Cellular Biology*, 19(2):1210–7, 1999.

[170] B.G. Turner and M.F. Summers. Structural biology of HIV1. *Journal of Molecular Biology*, 285(1):1–32, 1999.

[171] P. Uetz, Y.A. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S.V. Rajagopala, M. Roupelieva, D. Rose, E. Fossum, et al. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239, 2006.

[172] D. van Dijk, G. Ertaylan, C. Boucher, and P. Sloot. Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks. *BMC Systems Biology*, 4(1):96, 2010.

[173] T. Vavouri, J.I. Semple, R. Garcia-Verdugo, and B. Lehner. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, 138(1):198–208, 2009.

[174] C. Von Mering, L.J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research*, 35(Database issue):D358, 2007.

[175] S.L. Wong, L.V. Zhang, A.H.Y. Tong, Z. Li, D.S. Goldberg, O.D. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, et al. Combining biological networks to predict genetic interactions. *PNAS*, 101(44):15682, 2004.

[176] Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

[177] S. Wuchty, Z.N. Oltvai, and A.L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2):176–179, 2003.

[178] S. Yamada, M. Hatta, B.L. Staker, S. Watanabe, M. Imai, K. Shinya, Y. Sakai-Tagawa, M. Ito, M. Ozawa, T. Watanabe, et al. Biological and Structural Characterization of a Host-Adapting Amino Acid in Influenza Virus. *PLoS Pathogens*, 6(8), 2010.

[179] X. Yang and D. Gabuzda. Mitogen-activated protein kinase phosphorylates and regulates the HIV-1 Vif protein. *Journal of Biological Chemistry*, 273(45):29879–29887, 1998.

[180] X. Yang and D. Gabuzda. Regulation of human immunodeficiency virus type 1 infectivity by the ERK mitogen-activated protein kinase signaling pathway. *Journal of Virology*, 73(4):3460, 1999.

[181] M.L. Yeung, L. Houzet, V.S.R.K. Yedavalli, and K.T. Jeang. A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *Journal of Biological Chemistry*, 284(29):19463, 2009.

[182] L.S. Young and A.B. Rickinson. Epstein–Barr virus: 40 years on. *Nature Reviews Cancer*, 4(10):757–768, 2004.

[183] Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4), 2004.

[184] H. Zhou, M. Xu, Q. Huang, A. T. Gates, X. D. Zhang, J. C. Castle, E. Stec, M. Ferrer, B. Strulovici, and D. J. Hazuda. Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host & Microbe*, 4(5):495–504, 2008.