Prediction of Host,Virus Protein Protein Interactions

Oznur Tastan

CMU-LTI-11-009

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Judith Klein-Seetharaman (Co-chair, University of Pittsburgh, Carnegie Mellon University)
Jaime G. Carbonell (Co-chair, Carnegie Mellon University)
Ziv-Bar Joseph (Carnegie Mellon University)
Tom Smithgall (University of Pittsburgh)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

CARNEGIE MELLON UNIVERSITY
School of Computer Science

Prediction of Host,Virus Protein Protein Interactions

CMU-LTI-11-009

Oznur Tastan

Language Technologies Institute

School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15213

A dissertation submitted to Carnegie Mellon University

in conformity with the requirements for the degree of Doctor of Philosophy.

Thesis Committee:

Judith Klein-Seetharaman (Co-chair, University of Pittsburgh, Carnegie Mellon University)

Jaime G. Carbonell (Co-chair, Carnegie Mellon University)

Ziv-Bar Joseph (Carnegie Mellon University)

Tom Smithgall (University of Pittsburgh)

July 14, 2011

Abstract

New infectious viruses appear regularly, established ones fail to be eradicated, posing significant challenges to public health. Our lack of understanding of the intimate relationship between viruses and their hosts makes it difficult to develop effective therapies. Protein-protein interactions (PPIs) are key players in the cell generally and viruses exploit them for their purposes. Considerable progress has been made with HIV-1, the causative agent of AIDS, where experimental efforts have identified thousands of physical interactions and functional associations between the virus and the human host proteins. However, the complete and accurate repertoire of the physical interactome is still far from complete. Towards better defining the virus-host interactome, this dissertation complements experimental efforts by bridging different levels of biological information in a machine learning framework. Specifically, a wide array of genomic and proteomic data that could serve as direct and indirect feature evidence for virus, host PPIs was compiled. A supervised classification model was presented based on this data. A high quality label set was obtained by collecting experts' opinions on published interactions. A probabilistic framework was provided to estimate expert labeling accuracies and to obtain reliability scores for each interaction. Finally, to overcome data scarcity issues, we developed a multi-task learning strategy, where single tasks (learning the PPIs of each viral protein) shared parameters across different tasks based on their relatedness. The methods developed as part of this thesis can be easily extended to other host-virus systems as pertinent data become available. Numerous predictions of HIV-1, human interactions have subsequently been partially validated by experiments.

To my Mom and other women
who were deprived of educational opportunities

Acknowledgements

This work would not have been possible without the support of my collaborators, friends and family. Foremost, I thank my advisors, Profs. Judith Klein-Seetharaman and Jaime Carbonell, for providing me with a unique research environment where I could pursue my research interests. Judith's energy, enthusiasm and perseverance have always been inspiring. I am truly grateful to her for introducing me to diverse research questions and colleagues, guiding me to look at the problems through different angles and most importantly, being determined. At every turn, she generously provided her constant support, time and attention. I thank her family, Sridhar and Roshan, for joyful dinners and parties. I also thank Jaime for his great mentorship. I feel it is a rare opportunity to have had the chance to work with him. His suggestions were always constructive. He set up a great model of research vision and had amazing insight on not only which direction to go, but also which directions not to go.

Thank you to the members of my committee, Profs. Ziv-Bar Joseph and Tom Smithgall for their valuable suggestions, time and energy. I also thank HIV-1, Pittsburgh Center members and Profs. A.J. Rader, Hagai Meirovitch, with whom I had the chance to collaborate on interesting projects.

I thank the Carnegie Mellon School of Computer Science research community and staff members, who have created an environment where research is fun and students feel appreciated. I am grateful to my undergraduate school, Sabanci University, for not only their generous fellowships but also giving me the chance to experience the joy of research for the first time. Thanks to Profs. Osman U. Sezerman and Canan Atilgan, who let me dive in computational biology projects starting in my freshman year. Thanks to Prof. Kemal Oflazer, who encouraged me to apply to CMU.

I thank my colleagues and friends, who shared the pain, joy, laughter, and pride of my research over the years, sometimes from miles away: Berkem, Yase, Gokce, Gorkem, Andreas, Viji, Fernanda, Tankut, Jenn, Pinar, Volkan, Mehmet, Basak Alper, Elif, Ozge, Ibrahim, Emre Salman, Mohit, Lingyun, Fathima, Majid, Andy, Bilsay, Hasan, Tugba, Basak Isin, Umut, Alp, Soner, Gunes, Dilan, Emil, Andy, Pradipta, Henry, Amr, Betty, Yanjun, Madhavi, Sivaraman, Leman, Jose, Kalyan, Arpana, Naveena and many more. Thank you.

I couldn't have achieved anything without my family. When I stepped out of our home to sleep under the Science High School dormitory roof, I'm sure you did not foresee that this journey would take me to the other side of the world. Thanks a lot for respecting all of my decisions, bearing with the distance and loving me unconditionally. The toughest part was being away from all of you. I am also indebted to my fiancé's family, who were there with me at every peak and valley.

Finally, I am grateful to the series of events that brought me and my fiancé Osman Burak Okan together; with whom the universe runs in its order and the world becomes a much easier and more joyful place.

Contents

1 Introduction 1

1.1 Organization 4

2 Biological Background 5

2.1 The Central Dogma: DNA, RNA and Proteins 5

2.2 Protein-Protein Interactions 7

2.2.1 Methods for Detecting Protein-Protein Interactions Experimentally . 8

2.2.1.1 Small-Scale Protein-Protein Interaction Experiments 8

2.2.1.2 Large-Scale Protein-Protein Interaction Experiments 10

2.2.1.3 Experimental Methods Applied for Detecting Host-Virus Protein-Protein Interactions 12

2.2.2 RNA Interference Screens for Detecting Functional Associations . . 13

2.3 Viruses 14

2.3.1 Virus Structure 15

2.3.2 Viral Genome 15

2.3.3 Virus Life Cycle 17

2.3.4 Classification of Viruses 20

2.3.4.1 Human Immunodeficiency Virus-1 21

2.4 Genomewide RNA Interference Screens for Detecting Viral Host Factors . . 22

2.4.1 Genomewide HIV-1 Related RNA Interference Screens 23

3 Related Prior and Recent Work 25

3.1 Review of Prediction of Intra-Species Protein-Protein Interactions 25

3.1.1 Methods Based on Single Biological Evidence 26

3.1.1.1 Gene Fusion 26

3.1.1.2 Gene Neighbor and Gene Cluster Methods 27

3.1.1.3 Phylogenetic Profile Methods 27

3.1.1.4 Domain Profile Methods 28

3.1.1.5 Interolog Based Methods 29

3.1.1.6 Coevolution and Correlation of Phylogenetic Distances . . 29

3.1.1.7 Sequence-Structure Threading Based Methods 30

3.1.2 Methods Based on Multiple Sources of Biological Evidence 30

3.2 Review of Prediction of Inter-species Protein-Protein Interactions 34

3.2.1 Domain Profile Approach 34

3.2.2 Interolog Based Approach 35

3.2.3 Structural Similarity Based Approach 36

4 Thesis Overview 38

4.1 Predicting Virus-Host Interactions 39

4.2 Extended Model with New Feature Set 40
4.3 Refining Literature Curated Protein-Protein Interactions with Expert Opinions 41
4.4 Multi-Task Learning for Virus, Host PPI Prediction 43

5 Predicting the HIV-1,Human Protein Interactome 45

5.1 Overview 45
5.2 Methods 46
5.2.1 Problem Setting and Formulation 46
5.2.2 Classification 46
5.2.2.1 Decision Tree Classifier 47
5.2.2.2 Random Forest Classifier 47
5.2.3 Dataset 50
5.2.4 Features 53
5.2.4.1 Gene Ontology Features 55
5.2.4.2 Graph Topological Properties 57
5.2.4.3 ELM-Ligand Feature 57
5.2.4.4 Gene Expression Features 58
5.2.4.5 Tissue Expression Feature 59
5.2.4.6 Sequence Similarity Features 59
5.2.4.7 Posttranslational Modification Similarity to Neighbor . . . 60
5.2.4.8 HIV-1 Protein Type Features 60
5.2.5 Performance Evaluation 60

5.2.6 Assessing Features’ Predictive Power 61

5.2.7 GO Enrichment Analysis 61

5.3 Results and Discussion 61

5.3.1 Classifier Performance 61

5.3.2 Features’ Predictive Power 63

5.3.3 Comparison with Other Biological Information 66

5.3.4 Enriched Functions and Biological Processes 69

5.3.5 Colocalization of Selected Host Proteins with Vpr and Capsid . . . 70

5.4 Conclusions 74

5.5 External Collaborators 75

6 Extended Model with New Feature Set 76

6.1 Overview 76

6.2 Methods 77

6.2.1 Problem Setting and Formulation 77

6.2.2 New Features Included 78

6.2.2.1 Features Based on Interaction Data Derived from Immunoprecipitation Assay Coupled with Mass Spectrometry . . . 79

6.2.2.2 RNA Interference Screens Hits 80

6.2.2.3 Human Proteins Detected in HIV-1 Virions 84

6.2.2.4 Other Virus-Host Interactions 85

6.2.3 Performance Evaluation 85

6.2.4 Gene Ontology Enrichment Analysis 86

6.3 Results and Discussion . . . 86

6.3.1 Features' Contribution to the Performance . . . 87

6.3.2 Predicted Interactome . . . 90

6.3.3 An Independent Experimental Validation, SUMO2 Interaction with the Viral Protein Integrase . . . 90

6.3.4 Enriched Functions and Biological Processes . . . 91

6.3.5 Comparison of the First Model and the Old Model . . . 93

6.4 Conclusions . . . 93

7 Refining Literature Curated Protein-Protein Interactions with Expert Opinions 95

7.1 Overview . . . 95

7.2 Reliability of Literature Curated HIV-1, Human Protein-Protein Interactions 96

7.3 Approach . . . 98

7.4 Methods . . . 100

7.4.1 Problem Formulation . . . 100

7.4.2 Estimating Experts' Labeling Qualities . . . 101

7.4.3 Synthetic Data Experiments . . . 104

7.4.3.1 Experimental Set Up . . . 104

7.4.3.2 Baseline Estimators . . . 106

7.4.3.3 Evaluation . . . 106

7.4.4 Annotating HIV-1 Human Interaction Labels . . . 107

7.4.5 Refining the Literature Curated HIV-1, Human Interactome . . . 108

7.5 Results . . . 108

7.5.1 Synthetic Data Experiments . . . 108

7.5.2 Refining the Literature Curated HIV-1,Human Protein Interactome . 113

7.5.3 Model Trained with Expert Labels . . . 116

7.6 Conclusions . . . 118

8 Multi-Task Learning for Predicting Host,Virus Interactions 119

8.1 Overview . . . 119

8.2 Methods . . . 120

8.2.1 Data . . . 120

8.2.2 Multi-Task Random Forest Classifier . . . 121

8.2.3 Features . . . 123

8.2.4 Evaluation . . . 124

8.3 Results . . . 124

8.3.1 Performance of the Multi-Task Random Forest Classifier . . . 124

8.3.2 Model Trained with Multi-Task Random Forest Classifier . . . 126

8.3.3 Gene Ontology Term Enrichment Analysis . . . 126

8.4 Conclusion . . . 128

9 Conclusions and Future Work 129

9.1 Thesis Summary . . . 130

9.1.1 First Supervised Model for Predicting HIV-1, Host PPIs . . . 130

9.1.2 Improved Prediction Model for HIV-1, Host PPIs . . . 131

9.1.3 Obtaining High-Quality Interaction Labels . . . 132

9.1.4 Multi-Task Learning for Host,Virus PPI Prediction 133

9.2 Practical Usage of Predicted Virus-Host Interactions 134

9.2.1 Novel Predicted Pairs 134

9.2.2 Pairs with Prior Experimental Evidence 134

9.2.3 Utilizing the Information Related to Features 135

9.2.4 Identifying New Drug Targets 135

9.2.5 Alternative Pathways 135

9.2.6 Developing Broad Spectrum Antiviral Drugs 136

9.3 Future Work 137

9.3.1 Predicting Interactions in Other Host-Virus Systems 137

9.3.2 Predicting Functional Consequences of PPIs 138

9.3.3 Crowd-Sourcing for Refining PPI Databases 138

Bibliography **142**

List of Tables

2.1 Small-scale experimental methods for identifying protein-protein interactions . . . 9

2.2 Yeast two-hybrid studies focusing on virus-host interactome. . . . 13

2.3 Examples of cell receptors and virus proteins involved in attachment and fusion proteins . . . 18

2.4 Examples of human viruses and their characteristics . . . 20

2.5 Genome-wide RNA interference screens applied to detect host factors important for HIV-1 infection . . . 23

5.1 Group 1 and Group 2 keywords . . . 50

5.2 Number of interactions between HIV-1 proteins and human proteins reported in the NIAID HIV- 1, Human Interaction database . . . 51

5.3 Biological data sources . . . 53

5.4 Summary of features derived for HIV-1 and human protein-protein interaction prediction . . . 56

5.5 Gene expression data sets of HIV-1 infected and uninfected samples. . . . 58

5.6 Summary of results . . . 63

5.7 Number of predicted pairs at different precision-recall values . . . 66

5.8 Number of predicted pairs at different precision-recall values 68

5.9 Enriched GO molecular function terms in the set of human proteins that involve novel and Group 2 predicted interactions 69

5.10 Enriched GO molecular process terms in the set of human proteins that involve novel and Group 2 predicted interactions 70

5.11 Enriched GO cellular component terms in the set of human proteins that involve novel and Group 2 predicted interactions 70

5.12 Colocalization results for genes predicted to interact with capsid 73

5.13 Colocalization results for genes predicted to interact with vpr 74

6.1 Additional features derived for prediction of interactions between HIV-1 and human proteins. 78

6.2 Example of protein complexes enriched in the set of RNA interference detected human genes 83

6.3 Example of pathways enriched with RNAi genes. 84

6.4 Example of proteins targeted by multiple viruses. 85

6.5 Performance of the Random Forest model trained with extended feature set 86

6.7 Enriched GO molecular function terms in the set of human proteins that participate in the predicted HIV-1, host PPI network 91

6.8 Enriched GO biological process terms in the set of human proteins that participate in the predicted HIV-1, host PPI network 92

6.9 Enriched GO cellular component terms in the set of human proteins that participate in the predicted HIV-1, host PPI network 92

6.6 Performances of HIV-1, human Random Forest model using different feature sets 94

7.1 Notations used in this chapter 100

7.2 Two example synthetic data experiments for global annotation and subset annotation scenarios. 109

7.3 Performance of model trained with expert labeled data. 117

8.1 Number of expert labeled HIV-1 proteins, host PPIs. 120

8.2 Enriched GO biological function terms in the set of predicted human interaction partners. 127

8.3 Enriched GO molecular process terms in the set of human proteins that involve in the predictions 127

8.4 Enriched GO cellular component terms in the set of predicted human interaction partners. 128

List of Figures

2.1 Central dogma 6

2.2 HIV-1 genes and proteins 22

5.1 The HIV-1 human interactome defined by Group 1 interactions and Group 2 interactions 52

5.2 Schematic showing features that incorporate knowledge of the human protein interactome 54

5.3 Precision vs. recall curve of Random Forest classifiers trained on different subsets of features 62

5.4 Gini importance indices for each feature 64

5.5 Precision vs. recall curve when protein type features were removed 65

5.6 HIV-1 targets hub proteins 65

5.7 The HIV-1 human interactome defined by Group 1 interactions and Group 2 interactions 72

6.1 Histogram of scores derived from HIV-1, human mass spectrometry pull-down experiment 79

6.2 The subunits of the mediator complex identified by one or more genome-wide RNA inferences screens 81

6.3 Overlap among genome wide HIV-1 RNA inference screens 82

6.4 Comparison of the two models trained with the feature set 1 and the expanded feature set 87

6.5 Precision recall curves showing single feature contribution. 88

6.6 Precision recall curves of various models with different subset of features. 89

6.7 Comparison of the two prediction lists trained with the feature set 1 and the expanded feature set 93

7.1 Histogram of the number of publications supporting each interaction HIV-1, human protein interaction 97

7.2 Average mean squared errors in estimating expert labeling qualities 110

7.3 The average mean squared errors in estimating expert labeling qualities in subset and global annotation cases 111

7.4 Precision and recall rates of different labeling strategies 112

7.5 Descriptive statistics of expert annotated HIV-1, human protein interactions 113

7.6 Estimated HIV-1 expert labeling accuracies 114

7.7 HIV-1,human protein interaction network based on HIV-1 expert opinions 115

7.8 Precision recall curve of expert labeled trained models 116

8.1 Heatmap showing the predefined functional similarity of the viral proteins 123

8.2 Performance of multi-task Random Forest classifier. 125

List of Algorithms

1 Random Forest for classification 48

2 Set up of the synthetic data experiment to estimate expert labeling accuracies 105

3 Multi-task Random Forest Classifier 122

List of Symbols and Abbreviations

- $\mathcal{G}(\mathcal{V}, \mathcal{E})$: graph where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges
- k_v: degree of vertex v
- C_v: clustering coefficient of vertex v
- B_v: betweenness centrality of vertex v
- $x \in \mathcal{X} \in \mathbb{R}^d$: input example feature vector
- $h : \mathcal{X} \rightarrow \mathcal{Y}$: classifier function
- $y \in \mathcal{Y}$: target label
- $\hat{y} \in \mathcal{Y}$: estimated target label
- $p(x)$: data density distribution
- $p(y \mid x)$: posterior class label distribution
- $p(x \mid y)$: class label conditional density
- w: weight vector
- DAG: directed acyclic graph
- ROC: the receiver operating characteristic curve
- AUC: area under the curve
- MAP: mean average precision

- GO: gene ontology
- PPI: protein-protein interaction
- DNA: deoxyribonucleic acid
- RNA: ribonucleic acid
- RNAi: RNA interference
- siRNA: small interfering RNA or silencing RNA
- shRNA: small hairpin RNA or short hairpin RNA
- ptf: protein type features
- HIV-1: Human immunodeficiency virus subtype 1
- HAdV: Human adenovirus
- MPyV: Murine polyomavirus
- HPV: Human papillomavirus
- BPV: Bovine papillomavirus
- EBV: Epstein-Barr virus
- SV: Simian virus
- RSV: Rous sarcoma virus
- HCV: Hepatitis C virus
- ELM: eukaryotic linear motif
- TAP-MS: tandem affinity purification mass spectrometry
- Y2H: yeast two-hybrid
- E. coli: Escherichia coli
- genome: the complete set of genes of an organism

- proteome: the complete set of proteins expressed by a genome
- HPRD: Human Protein Reference Database

Chapter 1

Introduction

Infectious diseases caused by viruses continue to pose a major threat to public health. Currently, there are about 500 million people worldwide suffering from chronic infections, causing 3.5 million deaths annually [1]. While viruses such as human immunodeficiency virus (HIV-1), influenza virus or hepatitis C virus (HCV) fail to be eradicated, new viruses like bird or swine flu, SARS, West Nile, and Ebola emerge or cross the species barrier, infecting humans who have little or no immunity to these novel or reemerging viruses [2] . Moreover, since infections are not confined to single countries, widespread outbreaks develop easily; in the case of swine flu, 25,1401 individuals were infected, causing 2,545 deaths across the globe in only four months [3][1]. In addition to naturally occurring infectious diseases and their burden on societies, there is also the looming danger of the use of viruses as instruments of war and terror. The need for effective antiviral strategies, therefore, is pressing.

As the public demand for antiviral therapies grows, science is challenged by failed vaccine trials and emerging antiviral drug resistance[4, 5]. For many viruses such as HCV or HIV-1, currently there is no effective vaccine. Antiviral drug therapies notably succeeded in significantly improving the prognosis of infected individuals with access to treatment; however, they do suffer from important drawbacks. Apart from problems related to drug adherence, tolerability, and accessibility, there are also various issues that

[1] The statistics are as of August 23, 2009 and cases for reported and confirmed cases [3].

limit the success of current antiviral treatments. In particular, many RNA viruses [6] and some single stranded viruses [7] mutate rapidly and are able to confer resistance to the compounds targeting them in a very short period of time as a result of their error prone replication mechanism [5, 8]. Drug resistance raises the concern that even more challenging types of viruses are yet to come as viruses continue to evolve under the selective pressure of drugs [9]. Secondly, since current drugs are designed to target a specific viral enzyme, they have a very narrow spectrum and can only treat specific viral species or subtypes. Finally, viruses like Herpes or HIV-1 are able to lie dormant within cellular reservoirs [10, 11] so that both the immune system and drugs fail to purge the virus from the body completely; these dormant viruses are potentially able to replenish infection upon interruption of the treatment. The problems associated with current antiviral drugs and vaccine trials call for innovative approaches to developing new and improved antiviral treatments and preventative strategies.

Current antiviral strategies are limited as they exclusively focus on viral factors [12], which leads to a narrow drug spectrum and drug resistant viral strains. Viruses are obligate intracellular parasites and thus they are unable to replicate without the support of the host. Thus, an alternate strategy involves targeting the interactions of the viral factors with the cellular factors that are essential for the viruses, instead of targeting only the viral factors themselves [13–15]. A virus hijacks the cellular machinery so that it can successfully produce its progeny, while at the same time avoiding the host's immune system. For example, several enveloped viruses bud from the cell by making use of the host's endosomal sorting (ESCRT) complexes that normally regulate the formation of the multivesicular bodies of the endosomal pathway [16]. In addition, many viruses have mechanisms for disrupting the immune response against viral infection [17, 18]. Antiviral therapies that target these essential interactions are promising since cellular factors would not be expected to mutate under antiviral drug pressure [13]. Therefore, the virus may have difficulties in developing resistance against drugs targeting interactions between invariable cellular proteins. Moreover, the host cell includes many proven druggable targets such as cell surface receptors, protein kinases or nuclear receptors [14]. The testing of existing drugs currently used for unrelated diseases on viral infections could result in cost-effective compounds against viral diseases [15]. An additional challenge of targeting host factors is not disrupting (greatly) the normal cellular functions relying on the targeted host factors. Nonetheless, cellular functions are highly redundant as gene knockout studies in mice have shown [13].

Viruses exploit the cellular machinery through interactions between virus and host proteins. Proteins are key players in the cell, taking part in virtually all aspects of biological processes including catalyzing reactions of metabolism, transporting molecules, mediating signals from the exterior of a cell to the interior, carrying signals for transcription of genes, and forming structural entities. Proteins realize these functions in coordination with other proteins through *protein protein interactions (PPIs)*. Viral proteins, too, talk to the cell through interactions with the host's proteins. Therefore, to decipher the complex interplay between the host and the virus, interactions that occur between the viral and cellular proteins need to be identified. Advancing our understanding of this problem will provide the means for the rational design of novel intervention strategies for viral infections.

Experimental efforts to elucidate interactions between viruses and the host led to significant new insights about viruses such as HIV-1, influenza, and HCV. However, the map of interactions between the virus and the host is still far from complete for these and other host-pathogen systems, and active experimental efforts continue to identify these PPIs. In deciphering PPIs within a single organism (intra-species PPIs), computational methods have been instrumental [19], especially in the case of model organisms such as Baker's yeast and *E.coli*. Computational models accelerated experimental efforts to identify PPIs by suggesting hypotheses about novel interacting protein pairs or by stratifying the noisy high-throughput results [19]. Additionally, through the analysis of the set of interactions in a network framework, several genome-scale principles that might govern these networks have been identified [20].

While the literature on intra-species PPI prediction tasks is rich, the work on inter-species protein interaction prediction has been limited. A primary hurdle preventing progress in that area has been the scarcity of data sources. In recent years, the availability of genomic, proteomic, and phenotypic data increased drastically. By leveraging this accumulated information, I provide methods to computationally predict host-virus interactions.

1.1 Organization

The remaining chapters of this thesis are organized as follows:

- Chapter 2 provides a biological background on cells, viruses and protein-protein interactions.
- Chapter 3 summarizes the related literature on PPI prediction, focusing on inter-species PPI prediction.
- Chapter 4 states the open challenges and approaches pursued in this thesis. Thesis contributions are stated.
- Chapter 5 presents a supervised learning model for predicting binary physical interactions between HIV-1 and human proteins and provides experimental evidence for the validity of a subset of predictions.
- Chapter 6 reports an improved version of the model presented in Chapter 5, which expands the feature set with new biological information.
- In Chapter 7 the collection and analysis of curated expert opinions on HIV-1,human protein interactions are described. A probabilistic approach to estimating the reliability of reported interactions based on subjective expert opinions is presented.
- Chapter 8 presents a computational model for protein-protein interaction prediction based on a multi-task learning framework. Building upon the results from previous chapter an improved model is presented.
- Chapter 9 concludes the thesis and outlines possible future directions.

Chapter 2

Biological Background

2.1 The Central Dogma: DNA, RNA and Proteins

Deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins are key molecules of the cell. DNA stores the genetic instructions for how the cell develops and functions (with the exception of RNA viruses) [21]. These instructions contain information on how to synthesize other molecular components of cells, such as proteins and RNA molecules. DNA consists of long polymers of nucleotides with backbones made of sugars and phosphate groups. *Genes* reside on the DNA sequence (on RNA for RNA viruses) and are considered the hereditary unit of the living organism. A modern definition of 'gene' is 'a region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions' [22]. Genes hold the information necessary to build and maintain an organism's cells and pass genetic traits to offspring.

RNA is also one of the major macromolecules of the cell. Like DNA, it is a polymer of nucleotides; each nucleotide consists of a nucleobase, a ribose sugar, and a phosphate group. Unlike DNA, most RNA molecules is formed single-stranded in the cell. For RNA viruses, RNA stores the whole genetic information of the virus. For example, some viruses use RNA instead of DNA as their genetic material, and all organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins.

There are different types of RNA; messenger RNA (mRNA) carries information from DNA that is later translated into a protein sequence. Many RNAs do not code for protein [23]. The most prominent examples of non-coding RNAs are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation. There are also non-coding RNAs involved in gene regulation, RNA processing and other functions [24].

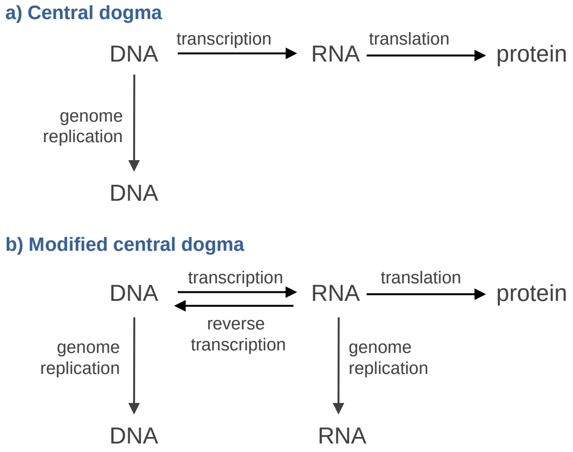


Figure 2.1: a) The central dogma of molecular biology in its first presented form. b) Modified central dogma which explains various models of virus transcription and genome replication.

Proteins are also polymers; their basic unit is an amino acid. There are 20 types of amino acids that differ in their chemical structures and thus physicochemical properties. Depending on the arrangement of these different amino acids in the protein sequence, the protein folds into a usually unique 3D structure. A fundamental principle of molecular biology is that protein structure governs protein function. The distinctive structures of proteins allow them to carry out a myriad of functions in the cell. Some examples of the variety of functions they implement are: i) catalyzing reactions of metabolism, ii) transporting molecules, iii) mediating signals from the exterior of a cell to the interior, iv) carrying signals for transcription of genes and v) forming structural entities. The term *genome* is used to refer to the complete set of genes of an organism, whereas the *proteome* is the entire complement of proteins expressed by a genome. Owing to the advent of sequencing technologies, the genome and proteome of many organisms, including humans, are now available [20, 25, 26]. Proteins seldom operate in isolation when carrying out these tasks; rather, they work in concert with other proteins. Major efforts are currently directed at discovering how these entities operate together to perform cellular functions.

Francis Crick coined the term ‘central dogma’ to describe the manner in which information flows from DNA to protein. The central dogma states that proteins are not directly synthesized from the DNA [27]; rather, they are first transcribed into an mRNA molecule (transcription) and then translated into protein sequences (translation), and the genetic information is transmitted from one generation to another through copying of the DNA (Figure 2.1 a). Increasing understanding of viral genome replication necessitated modifications to the central dogma in 1970 [28]. Many viruses have RNA genomes that are copied to RNA, and some viruses copy from RNA to DNA (Figure 2.1 b).

2.2 Protein-Protein Interactions

The physical contact made by the proteins with themselves and each other are referred to as ‘protein protein interactions’ (PPIs). While a physical contact is the prerequisite for the most narrow definition, the phrase ‘protein-protein interaction’ has been used as a fairly broad term in the literature to refer to a wide range of relationships among proteins, from direct physical interaction to functional associations. For example, proteins might interact through direct physical contact, where the two proteins are bound to each other. Such an interaction often results in one protein modifying another protein via this interaction, i.e. a kinase protein adding a phosphate to the target protein. In *functional associations*, in contrast, two proteins might cooperate to carry out a given task without actually (or necessarily) engaging in physical contact. This could occur through being part of *protein complexes*, which are units where two or more proteins collaborate together to exert a function. In this case, two or more proteins may indirectly interact through physical contact with a third protein. Another form of functional association is participation in the same pathway. A *pathway* is a cascade of interactions between proteins and other molecules that, when activated, ultimately change some aspect of cell behavior [29]. Throughout the thesis, it will be made clear what type of interaction I am referring to in each context.

PPIs can be classified in other overlapping ways as well. For example, reference to homo- versus hetero-oligomeric interactions specifies whether the proteins participating in an interaction are identical or not. PPIs can also be grouped as stable versus transient interactions. Stable interactions usually construct macromolecular structures such as the cytoskeleton or the mitotic spindle. Transient PPIs, on the other hand, are usually involved in the regulation of fundamental cellular processes such as protein modification, transport, signaling, cell cycling, etc. Finally, PPIs can be grouped as strong vs. weak interactions based on their binding affinity. Interactions between proteins within a cell are tightly regulated at multiple levels through expression, post-translational modifications, or ligand binding [30], enhancing the complexity of the PPI dramatically.

Since the set of interactions of a protein largely determines, identifying PPIs within an organism and between organisms is essential to uncovering how cellular processes are executed. Therefore, the development of methods to detect and characterize PPIs has been a major theme of functional genomics and proteomics efforts. The data used in this dissertation make use of experimentally derived PPIs. The motivation of predicting PPIs comes from the fact that there is no single cost-effective reliable experimental technique that allows high-throughput identification of PPIs. Therefore, in the following section experimental methods for detecting PPIs will be introduced.

2.2.1 Methods for Detecting Protein-Protein Interactions Experimentally

In order to identify PPIs experimentally, a wide array of techniques are available including genetic, biochemical, and physical techniques (reviewed in [31]). These can be broadly classified as small-scale and large-scale experiments based on the number of proteins studied at a time.

2.2.1.1 Small-Scale Protein-Protein Interaction Experiments

Small-scale experiments interrogate a small number of PPIs ($\leq$ 10) at a time in a hypothesis driven approach as opposed to screening all interactions in a high-throughput fashion. Small-scale experiments are often very labor-intensive and time-consuming. These techniques include biochemical, genetic and biophysical methods such as co-

Method	Description
Co-immunoprecipitation assays	The protein of interest is tagged by an antibody to which it binds in a sample solution; the bound proteins co-precipitate with the bait. Low affinity or transient interactions are hard to detect.
Pull-down assays	Similar to co-immunoprecipitation assay, except a bait protein is used instead of an antibody to purify any proteins in a lysate that bind to the bait. Ideal for studying interactions for which no antibody is available for co-immunoprecipitation. Ideal for strong or stable interactions.
Crosslinking	The proteins of interest are chemically cross-linked to each other, so that interaction is fixed before the isolation of interacting proteins by a complementary technique.
Colocalization	Checks whether two proteins are located in the same area or very near each other in the cell. Fluorescence Resonance Energy Transfer is one way of achieving this proximity check. Used usually as a primary screen, followed by more detailed studies.
Fluorescence Resonance Energy Transfer (FRET)	Two proteins are tagged with different fluorophores and expressed in a cell. If those two proteins interact, the two labels come in close proximity and a detectable loss of excitation energy is measured.
Surface Plasmon Resonance (SPR)	Infers whether a protein interacts with the bait protein, based on small changes in laser light reflected from a film coated with the bait protein.
Far Western	Protein samples of interest are immobilized on a membrane and probed with a putative interaction partner.
X-ray crystallography	Interacting proteins are co-crystallized and the structure is resolved through their X-ray diffraction patterns, characterizing the interaction on an atomic level.
NMR	The interacting proteins are resolved in solution through Nuclear Magnetic Resonance. Characterizes the protein interaction in atomic detail.

Table 2.1: Examples of small-scale experimental methods for identifying and characterizing protein-protein interactions.

immunoprecipitation, fluorescence resonance energy transfer (FRET) studies, Nuclear Magnetic Resonance (NMR), and X-ray crystallography. Some of these techniques are listed in table Table 2.1. Each experimental technique has certain limitations and powers, which depend on the properties of proteins under study and the nature of their interactions. For example, many techniques fail to capture transient interactions since the interaction may dissociate during the study. In such cases, crosslinking experiments are powerful because they trap the interaction in place. However, crosslinking may increase the false positive rate. For some methods, a positive result would indicate binary interactions, whereas in others a positive result might indicate both direct and/or indirect interaction; e.g. co-immunoprecipation experiments. On the other hand, very labor intensive techniques such as NMR and X-ray crystallography can characterize the details of the interaction with an atomic resolution.

2.2.1.2 Large-Scale Protein-Protein Interaction Experiments

In contrast to small-scale experiments, large-scale techniques enable high-throughput screening of PPIs. Widely employed large-scale PPI detection methods include yeast two-hybrid (Y2H) [32] and related assays such as the split ubiquitin system [33], affinity purification (AP) [34], usually coupled with mass spectroscopy (MS) [35], and DNA and protein microarrays [36]. The drawback of these methods is their high false positive and negative rates [37, 38]. Below the two most popular large scale methods, Y2H and TAP-MS, are described:

Yeast Two-Hybrid Assay: Y2H system [32] is widely used to discover PPIs *in vivo*. The Y2H method is based on the principle that eukaryotic transcription factors' activation and binding domains can function in close proximity even though they are split into two domains. Relying on this principle to the Y2H technique utilizes activation of a downstream reporter gene by the binding of a transcription factor onto an upstream sequence. In the Y2H system, the protein of interest, referred to as 'bait', is typically fused to a DNA-binding domain (DBD). The other protein of interest, referred as 'prey', is fused to a transcription-activating domain (TAD). If the bait and the prey interact, the DNA binding domain and the activation domain come into proximity of each other and restore the function of the transcription factor. As a result, the transcription of the reporter gene is triggered. The reporter gene enables growth on specific media or a color reaction. The Y2H technique can be used both on a large scale and for a small set of interactions.

Although Y2H system is regarded as one of the most powerful methods of identifying PPIs, it has also been criticized for its high false positive and false negative rates

[39]. Screens conducted in different laboratories often do not overlap. This lack of reproducibility is even more pronounced when Y2H data is compared to datasets derived from other large scale techniques, such as affinity purification/mass spectrometry (AP/MS) experiments [40]. The difference is usually attributable to use of different vectors, strains, or reporter genes. Another inherent drawback of the method is that the reaction takes place in the nucleus; therefore, the proteins under study are typically not in their native compartment. High-throughput Y2H screening has been applied to several species to detect pairwise direct interactions within the entire proteome (all possible proteins) of a given organism [43, 44, 44–50].

Affinity Purifications-Mass Spectrometry: Affinity purification followed by mass spectrometry (AP-MS) identification is a powerful method of studying novel interactions [50, 51]. In contrast to yeast two-hybrid experiments, which only reveal pairwise interactions, AP-MS experiments allow identification of PPIs in a complex [52]. The method involves biochemical isolation of protein complexes using an inherent interaction (affinity) and subsequent identification of their constituting proteins using mass spectrometry [53, 54]. One of the molecules is immobilized on a solid support, and the interacting molecule is purified along with associated proteins. There are many different affinity reagents such as antibodies or other recombinant proteins, which may be epitope-tagged. Co-immunoprecipitation mass spectrometry experiment (IP-MS) is an AP-MS method, in which antibodies are used in the isolation step. The protein complex is captured from cell lysates by an immobilized antibody, which specifically recognizes an epitope of one component of the complex. The retrieved complex is washed to remove unspecifically bound proteins. Tandem affinity purification (TAP-MS) experiment is another AP-MS technique which allows high-throughput characterization of complexes. In TAP-MS experiments two tags instead of one is utilized sequentially. The protein of interest, *bait* protein, is first tagged via attachment of a purification tag to the polypeptide; next, the bait protein is expressed inside the cell. Using the tag of the bait protein, the complex is purified from a cell lysate via affinity chromatography to identify prey proteins forming protein complexes with the bait protein.

Following the isolation step, the identities of the prey proteins are introduced into the mass spectrometer to separate them according to mass (detected as mass-to-charge ratios). Peptides of a fixed size are selected and are broken into fragments. The resulting fragments are analyzed, which produces a peptide 'fragmentation profile'. Using this profile, proteins are identified by searching the resulting peptide mass fingerprints through sequence databases or via statistical classifiers. Ideally, these proteins would constitute the entire complex encompassing all proteins interacting with the bait protein but both the fragmentation peptide detection and protein spectrum reconstruction generate some noise. Although this experiment identifies the components of the protein complex, it typically does not provide information about interaction topology. AP-MS allows the detection of complexes in physiological settings; however, it may miss complexes that are not present under the experimental conditions. Additionally, tagging and purification may dissociate the complexes, so if they are weakly associated they may escape detection. This method therefore suffers from a high false negative rate.

The success of a PPI detection experiment, apart from its inherit limitations and powers, also depends on many other experimental factors, such as the ability to mimic the interaction conditions (buffer composition, pH, cofactor requirements), the concentration of the proteins and other requirements of the proteins, including post-translational modifications. To date, there is no reliable, cost-effective technique that can work on a large scale with high sensitivity and specificity; therefore, computational methods are used to stratify these studies.

2.2.1.3 Experimental Methods Applied for Detecting Host-Virus Protein-Protein Interactions

Majority of the host-virus interaction identification studies have been conducted via small-scale experiments (see Section 2.2.1.1). As an example, PPIs between HIV-1 and human proteins reported in the literature were cataloged and made available in the NIAID HIV-1,Human protein interaction database (NIAID database) [134, 135]. The database includes more than 2500 interactions, all of which are results of small-scale experiments. Large-scale experiments have been started to been applied to detecting virus-host interactions (see Table 2.2). Calderwood et al. [41] applied Y2H screen to the Epstein-Barr virus(EBV) and human system, which revealed 173 PPIs between 112 human and 40 EBV proteins. In another study, Y2H screen was applied to characterize HCV-human PPI network; this screen detected 314 PPIs between 278 human proteins and 11 HCV proteins [42]. Krogan et al. (unpublished) performed an affinity purification screen coupled with mass-spectrometry.

Virus	**Number of Interactions Detected**	**Number of Viral Proteins**	**Number of Human Proteins**	**Reference**
Epstein-Barr virus	173	40	112	[41]
Hepatitis C virus	314	11	278	[42]

Table 2.2: Yeast two-hybrid screens applied to detect virus-host protein-protein interactions.

2.2.2 RNA Interference Screens for Detecting Functional Associations

In addition to protein-protein identification methods where physical interactions direct or in a complex are detected, functional screens are available to identify functional associations. RNA interference (RNAi) or RNA silencing is a post-transcriptional gene silencing mechanism induced by double-stranded RNAs (dsRNAs), which deplete the complementary mRNAs in a cell in a sequence-specific manner. The RNAi silencing mechanism was first discovered in *Caenorhabditis elegans* [55] and has been reported to be endogenously present in such diverse organisms as plants, fungi and mammals [56]. It has been shown that the RNAi pathway undertakes fundamental regulatory roles for gene activity and structure [57, 58]. RNAi pathways are triggered by dsRNA molecules that are complementary to the target mRNA. A protein complex containing Dicer [59] cleaves the long dsRNAs into small interfering RNAs; these small interfering RNAs are also known as short inferring RNAs (siRNAs), and are 20-24 nucleotides long. siRNAs in turn are incorporated into the RNA-induced silencing complex (RISC). Following its assembly, the RISC targets the complementary mRNAs for degradation.

Upon the discovery that synthetic dsRNA exogenously introduced into eukaryotic cells can reduce the expression of a gene in a sequence specific manner, this process became a powerful technique for studying gene function [60]. Either short double-stranded, in vitro-synthesized siRNAs or short hairpin RNAs (shRNAs), expressed stably in cells from specialized DNA-based vectors, are delivered into the cells. Now, RNAi libraries covering full genomes are available and high-throughput analysis of each gene in cells is possible [56, 61]. By introducing si/shRNAs into the cell to silence target gene mRNAs, genome-wide RNAi screens enable interrogating the effect of silencing each gene in a phenotype of interest.

Current RNAi technologies have a number of limitations. Not all RNAi sequences are equally effective and potent; therefore, the design of si/shRNA libraries is critical. Most libraries contain pools of multiple independent siRNAs, usually three or four, for use when targeting one mRNA to increase the likelihood of a successful knockdown [60]. After an initial hit list is generated, computational approaches are applied to reduce the number of false positives. For an RNAi screen to be successful, biologically meaningful reduction in the mRNA levels with the siRNAs used must be achieved. Reporting assays are important when validating whether any given siRNA silences the targeted gene effectively. As siRNAs exert their effects at the mRNA level, the preferred assay for siRNA validation is the one that effectively monitors mRNA levels. One of the most sensitive assays for siRNA validation relies on qRT-PCR, which measures the target transcript levels in gene specific siRNA-treated and negative control cells.

Even when the mRNA is efficiently depleted, if the gene product is sufficiently stable such that it decreases too slowly to be monitored by the reporter phenotype, the gene might not be identified successfully [62]. Furthermore, siRNAs deplete a single gene product at a time, so a genome wide screening will not identify genes whose function can be carried out by other genes. Knocking down unintended target genes, referred to as off target effects, is another known problem with the RNAi screens [63]. Finally, the toxicity of a resulting knock-down is also an issue; the genes can only be knocked down to a level which is not toxic to the cell [62]. Regardless of these shortcomings, RNAi screens are of enormous interest and are widely used as primary screens.

2.3 Viruses

Viruses are small obligate intracellular parasites and are inert outside the host cell [64]. A fully assembled infectious virus is called a *virion*, and is composed of a nucleic acid (RNA or DNA) and proteins encoded by this genome. The nucleic acid of the virus contains all the information needed to produce new viruses by interacting with host cells. This information includes how to make new viral particles and accessory proteins, as well as how to redirect the host cell machinery for reproduction of viruses. The structure of the virion is an indicator of the viral requirements during its replication cycle.

2.3.1 Virus Structure

Viruses range in size from less than 100 nanometers to several hundred nanometers in diameter [65]. They display a wide diversity of shapes and sizes, known as morphologies; these varied structures reflect the efficiency and stability constraints faced by the virus. The viral genome is packaged inside a protein coat along with some viral proteins to avoid degradation by nucleases. This protective coat is called the *capsid*. The capsids are made of multiple copies of a limited number of protein species, which assemble in large numbers to form a continuous three-dimensional structure. This three-dimensional structure can be arranged such that the proteins are wrapped around a helical filament of nucleic acid or can take on an icosahedral morphology, a shape characteristic of the nucleocapsids of many 'spherical' viruses.

Many viruses encode relatively few structural protein species, as well as a few accessory proteins that participate in the replication of the viral genome [65, 66]. There are some viruses with proteins, most of which participate in replication, but are not packaged into the virion. Additionally, the capsids of some virus types are surrounded by extra envelope; this envelope is a protein-rich lipid membrane bilayer acquired in part from the host cell during budding. Several classes of proteins are associated with virus envelopes. Virus encoded *matrix* proteins link the envelope to the core of the particle; glycoproteins are responsible for receptor recognition and binding. Thus, in addition to virus-specified envelope proteins, viruses may also carry some host cell proteins as integral constituents of the viral envelope.

2.3.2 Viral Genome

Unlike all living organisms, whose genetic material is composed of double stranded DNA molecules, viral genomes can be made up of either RNA or DNA, which may be single stranded (ss) or double stranded (ds), and either linear or circular. The entire genome may be composed of a single nucleic acid molecule or several nucleic acid segments. The different types of genome result in different replication strategies [65].

In the case of the DNA viruses, the nucleic acid is usually linear, though some may have circular DNA. DsDNA serves as a template for the viral mRNA and for self-transcript. The capsid is made of two or three structural proteins; additionally, there are five to six nonstructural proteins encoded that play roles in virus transcription, DNA replication and cell transformation.

RNA viruses constitute the largest group of all viruses. In replication, the viral RNA is first transcribed into the DNA. A key property of these viruses is that the viral enzymes that are involved in this transcription are error prone and, due to a lack of proofreading mechanisms, these viruses' genomes mutate at a higher rate than the DNA viruses [65]. This high mutation rate gives the virus the capacity to adapt to new hosts and evade host defense mechanisms.

The RNA strand of a single-stranded genome may be either a sense strand (plus strand) or an antisense strand (minus strand). Sense RNA can function as mRNA, whereas antisense is complementary to the sense strand and cannot function as mRNA during protein translation. RNA viruses occur in four distinct groups depending on the number of RNA strands and whether the virus carries a sense strand or antisense strand:

- Viruses with a genome that consists of single-stranded antisense RNA; that is, RNA that is the complement of the message sense. This is also called negative-stranded RNA. Measles and Ebola viruses are examples of this type of virus.

- Viruses with a genome that consists of single-stranded sense RNA; that is, the RNA has message sense and can act as mRNA. This is also called positive-stranded RNA Poliovirus is an example of this type of virus.

- Viruses with a genome that consists of several pieces of double-stranded RNA; an example is reovirus.

- Retroviruses' genomes comprise two identical plus-sense ssRNA molecules. Their RNA (also single-stranded) is copied by viral reverse transcriptase into a DNA genome within the host cell; HIV-1 is a member of this group. Retroviruses contain two envelope proteins encoded by the *env* gene, 4-6 core proteins encoded by *gag* gene and 3 non-structural functional proteins specified by the *pol* gene. The reverse transcriptase transcribes the viral ssRNA into double-stranded circular proviral DNA. This DNA, mediated by the viral integrase, is inserted into the DNA of the host cell to make possible the subsequent transcription of the sense strands [67].

2.3.3 Virus Life Cycle

Although the precise details vary among individual viruses, all viruses go through a common sequence of event when replicated [65]. These steps include:

Attachment and viral entry: The viral infection of a host cell starts with attachment of the virus onto the surface of a susceptible cell by means of surface viral proteins, which interact with receptor structures on the host cell. These receptors take part in the normal functioning of the cell, but the viruses have evolved means to take advantage of them. The receptor molecules on the target cell are usually proteins; carbohydrates or occasionally lipids may also be used. The virus-receptor interaction is often specific to a particular virus but can also be common across a family of viruses. Some examples of cell receptors are given in Table 2.3. There are different kinds of receptor molecules: low affinity receptors, primary receptors and co-receptors. The attachment of the virus onto the surface of the host cell serves to overcome any repulsive forces that may exist between the virus and the cell, and also facilitates viral entry [66]. The attachment is followed by the internalization of the virus into the cell. Different types of viruses achieve this in different manners. The enveloped viruses either fuse with the plasma membrane, releasing the contents of the virion directly into the cell cytoplasm, or the enveloped viruses enter via endosomes at the cell surface. By contrast, non-enveloped viruses penetrate into the cell directly or are taken up into the endosomes; the endosome is later destroyed.

Uncoating: In order for replication to take place, the viral genome has to be accessible. Therefore, the virion is uncoated rapidly after cell penetration. This process includes either dissociation or partial degradation of the particles, alone or with the aid of cellular enzymes. In viruses such as papillomavirus or herpesvirus, the nucleocapsid is transported to the nuclear pore, where the viral DNA is released directly into the nucleus. In the case of some viruses, i.e. reoviruses, the capsid is only partially dissociated, and the

Virus	**Host receptor**	**Virus protein(s) involved in**	
		attachment to receptor	**fusion**
Naked viruses			
≈ 90% of human rhinoviruses	ICAM-1	VP1 + VP3	
≈ 10% of human rhinoviruses	Low-density lipoprotein receptors	VP1	
Poliovirus	CD155	VP1	
Enveloped viruses			
HIV-1	CD4	gp120	gp41
Influenza viruses A& B	Sialic-acid-containing glycoproteins	Haemagglutinin	Haemagglutinin
Measles virus	Signaling lymphocyte activation molecule (CD150)	Haemagglutinin	Fusion

Table 2.3: Examples of cell receptors, virus proteins involved in attachment and fusion for enveloped viruses. The table is adapted from Table 5.1 of [66].

viral genome expresses its functions without being fully released from the capsid. There are also viruses for which penetration and uncoating take place simultaneously. After uncoating, the viral genome is available as a naked nucleic acid or as a nucleoprotein complex.

Synthesis of viral nucleic acid and proteins: Once the virus genome has been made available, it needs to replicate and viral proteins need to be synthesized. All viruses use the protein synthesis machinery to translate viral mRNAs. Unlike the host's genetic material, which is encoded in double stranded DNA, the viral genome can take different forms, as described in Section 2.3.2. Depending on the viral genome type, the replication and transcription of viral mRNAs take place in different ways. In positive-stranded RNA viruses, viral RNA is translated directly from the viral genome. In negative-stranded RNA viruses, a positive-stranded RNA is produced from the original negative strand, and then newly produced copies are translated directly to viral proteins. In the case of the retroviruses, this process is more complex. A DNA-RNA hybrid form is first produced from its RNA genome via the virus-associated enzyme reverse transcriptase. The RNA molecule is then digested and replaced by a DNA copy, which results in a dsDNA molecule. The dsDNA molecule is integrated into the host's genome by the virus-encoded integrase enzyme. The virus finally replicates as part of the host cell's

DNA. Viral mRNAs are transcribed from this proviral DNA. On the other hand, most DNA viruses produce mRNA transcripts through a host-cell enzyme, DNA-dependent RNA polymerase II. An exception is the poxvirus, which carries the appropriate enzyme into the cell. Once the viral mRNAs are transcribed, they are translated into proteins using the host protein synthesis machinery. The replication of the genome is also different for different virus types and takes place in different locations. In the case of DNA viruses, with the exception of poxviruses, the viral genome is replicated in cytoplasm. A positive-stranded RNA virus, whose genome can act as the mRNA, may not need to enter the nucleus of the cell. With the positive-stranded RNA viruses, a virus-coded replicase is translated directly from the viral genome. In the case of negative-stranded RNA viruses, the virus carries the replicase itself. Either way, the RNA replicase synthesizes a complementary RNA strand that serves as a template for new rounds of viral RNA synthesis. These RNA duplexes are unstable and occur only as transient 'replicative intermediates'. In this manner, the RNA virus replicates. The process is rapid, with production of tens of thousands of new viral genomes produced in few hours; the results are error-prone because of the low fidelity of the reverse transcriptase and RNA transcriptase, as well as the absence of a proof-reading mechanism. This error-prone process results in new viral genomes with several mutations. In fact, all RNA viruses are thought to exist as mixtures of with slightly different genetic compositions, called *quasi-species*.

Assembly: Once the new viral genome and viral proteins are replicated, they are assembled to form the next generation of viruses. Some viruses assemble completely in the cytoplasm, whereas for other viruses the assembly takes place predominantly in the nucleus. The assembly of virions of many viruses involves the construction of a protein shell, known as procapsid. The procapsid contains the viral genome and protects it from the environment. During or after the assembly, the capsid may undergo modification to form the mature capsid. For some viruses modification of the virus involves cleavage of one or more of the viral structural proteins. On the hand, it also needs to be able to release the genomic content during infection, so the structure at the same time needs to be unstable when needed. For some viruses, assembly of the structure of the viral particle and budding occur simultaneously, whereas in others a preformed core pushes buds through the membrane.

Release: In the final state, new virus particles are released and begin to seek new potential host cells to start the life cycle anew. Envelope viruses do not necessarily kill the cell, but instead bud from the cell; by contrast, non-envelope viruses break the cell and cause cell death. The particles produced within the cell may require further processing to become infectious; such maturation may occur before or after release.

In all the phases of the virus life cycle, host-viral PPIs are essential, and the host machineries required depend on specific viral needs.

2.3.4 Classification of Viruses

Virus Family	**Examples**	**Enveloped or Naked Virion**	**Capsid Symmetry**	**Strand type***
		DNA viruses		
Adenoviridae	Adenovirus	Naked	Icosahedral	ds
Papillomaviridae	Papillomavirus	Naked	Icosahedral	ds, circular
Herpesviridae	Herpes simplex virus, cytomegalovirus, Epstein-Barr virus	Enveloped	Icosahedral	ds
Hepadnaviridae	Hepatitis B virus	Enveloped	Icosahedral	circular, partially ds
		RNA viruses		
Flaviviridae	Hepatitis C virus, yellow fever virus	Enveloped	Icosahedral	ss
Orthomyxoviridae	Influenzavirus A,Influenzavirus B, Influenzavirus C	Enveloped	Helical	ss (-)
Retroviridae	Human immunodeficiency virus 1, Human T-lympotropic virus I	Envoloped	Icosohedral	ss (w/DNA intermediate)

Table 2.4: Examples of human viruses and their characteristics. * ds: double stranded, ss: single stranded, (+): positive sense, (-): negative sense

.

More than 80 families of more than 30,000 different virus isolates are known today [68]. The International Committee on Taxonomy of Viruses (ICTV) assigns a virus into a taxonomic group by considering a range of characteristics. These include host range (eukaryote or prokaryote, animal, plant, etc.), morphological features of the virion (enveloped, shape of capsid or nucleocapsid, etc.), and the nature of the genome nucleic acid (DNA or RNA, single stranded or double stranded, positive or negative sense, etc.). Within these parameters, additional features are considered. Table 2.4 lists some viruses and their classifications.

2.3.4.1 Human Immunodeficiency Virus-1

HIV-1 is the etiologic agent of acquired immune deficiency syndrome (AIDS), and continues to be a major health threat [9, 69]. The number of AIDS-related deaths was approximately two million in 2007 alone; an estimated 33 million people worldwide are infected [70]. Use of antiretroviral therapy has prolonged patients' lives, but cellular latency and drug resistance problems remain. In addition to its medical importance, studying the HIV-1, human system computationally is motivated by the fact that it is a virus-host system, where data is relatively more abundant than other virus-host systems.

HIV-1 is a retrovirus, so it contains two copies of a single stranded RNA genome and is a member of the *Lentivirus*. The RNA genome includes nine genes (*gag, pol, and env, tat, rev, nef, vif, vpr and vpu*) encoding 15 proteins (19 including the prototypically cleaved forms) (see Figure 2.2). Three of these genes, *gag*, *pol* and *env*, code for major structural proteins common to all retroviruses. *gag* provides proteins to create the basic structure of the virus such as matrix protein (MA, p17), capsid protein (CA, p24), spacer peptide 1 (p2); nucleocapsid protein (p7), spacer peptide 2 (p1) and p6. *pol*, on the other hand, codes for viral enzymes reverse transcriptase, integrase, and HIV-1 protease. The *env* gene codes for gp160, the precursor to gp120 and gp41, all of which are proteins part of the viral envelope that enable the virus to attach to and fuse with target cells. There are also two regulatory proteins (tat and rev) and three accessory proteins nef, vif, vpr, vpu. None of the HIV-1 accessory proteins display enzymatic activity, but they are important in altering the cellular pathways of the host cell. In order to complete its replication cycle and escape the immune system at the same time, HIV-1 protein interact extensively with cellular factors.

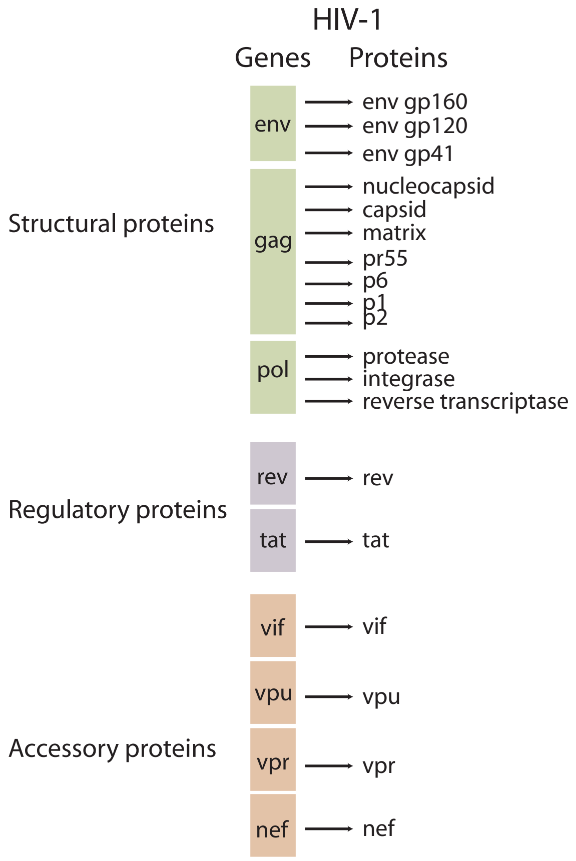


Figure 2.2: The 9 genes of the HIV-1 RNA genome and the proteins encoded by each. The different colors of the genes indicate different functional classes of proteins.

2.4 Genomewide RNA Interference Screens for Detecting Viral Host Factors

RNA technologies have great potential for dissecting gene functions such as the role of host genes in viral infections. In RNAi screens applied to viruses, host factors required for the viral infection are investigated (see Section 2.2.2). Genome-scale RNAi screening is used to identify host cell factors that promote or inhibit infection when the host gene is silenced [71]. The screens call a set of host genes that are potentially required for successful viral infection. The resulting set of hit genes contains a list of potential genes required for a successful viral infection. However, they are not necessarily direct interactors of the viral proteins, but rather can be cellular factors that are indirectly affecting the viral-host interactions. A number of genome wide RNAi screens have been applied to human viruses including HIV [72–74], HCV [75, 76], West Nile Virus [77] and influenza virus [78]. The results of the RNAi screens applied to HIV-1 are summarized below.

2.4.1 Genomewide HIV-1 Related RNA Interference Screens

RNAi screen	Number of genes called	Description	Ref.
siRNA Brass	281	Genome-wide siRNA screen in HeLa cells	[72]
siRNA König	295	Genome-wide siRNA screen for early stage of replication in 293 cells	[74]
siRNA Zhou	291	Genome-wide siRNA screen in 293T cells	[73]
shRNA Yeung	224	Genome-wide shRNA screen in Jurkat T-cells	[79]

Table 2.5: Genome-wide RNA interference screens applied to detect host factors important for HIV-1 infection.

Four HIV-1 related genome-wide RNAi screens have been conducted to identify host factors required for viral infection, summarized in Table 2.5. In these screens, each host gene was silenced and its effect on HIV-1 infection was measured. The hits generated by these screens are potentially critical host factors for viral replication. Of the four screens, the Yeung et al. study [79] employed a short hairpin RNA (shRNA) library cloned in a retroviral vector to knock out RNAs in Jurkat T-cells, while the other three screens employed presynthesized siRNA libraries in HeLa or 293T cells. HeLa and 293T cells are not natural HIV-1 target cells; however, they are highly efficient model cells for siRNA transfection. On the other hand, the Jurkat T-cells are are better models but cannot be efficiently transfected by siRNA [79]. The König et al. study [74] only examined the steps of uncoating through viral gene expression, while the other three studies [72, 73, 79] set out to detect host factors required for the whole replication process.

Although these screens were aimed at identifying the complete set of host factors (with the exception of the König et al. study), the resulting gene sets lack overlap. There is no single gene that is called for in all four RNAi screens, and there are only three genes

(RELA, MED6, and MED7) that are identified by three of the four screens. There are 36 genes identified by two screens, and 1010 genes are called by only one screen. The lack of overlap could be partially attributed to differences in experimental design i.e. RNAi libraries, cell types, reporter assays used or bioinformatics methods applied in post-processing the initial hit lists [62]. Additionally, the limitations of RNAi experiments, reviewed in Section 2.2.2, generally result in false positives and false negatives.

Chapter 3

Related Prior and Recent Work

Computational methods have the potential to accelerate experimental efforts in identifying PPIs; improving the coverage, accuracy, and efficiency of PPI detection. An array of computational methods has been proposed for predicting direct physical interactions, and complex or pathway memberships. A large fraction of these studies are designed for detecting PPIs within a single organism, which we refer to as the *intra-species prediction task*, whereas little work has been devoted to detecting the inter-species case PPIs between pathogen and host organisms, referred to as the ‘inter-species’ prediction task. In this chapter, we will first review the methods proposed for the intra-species PPI prediction task, and then review the work on predicting inter-species PPIs.

3.1 Review of Prediction of Intra-Species Protein-Protein Interactions

Computational approaches for predicting PPIs can be divided into two groups: i) those that predict novel PPIs based on a single biological piece of information and ii) those that integrate multiple pieces of information to make their predictions. The first group comprises an earlier set of work; these methods depend on a single piece of genomic or proteomic biological information that is used as evidence for possible interactions. In this approach, the resulting list of interactions is typically pruned using additional information, such as cellular location of proteins. Although this second step reduces the false positive rate, it does not have an effect on the frequency of false negatives. Therefore, the additional information sources are not exploited to their full extent for prediction of novel interactions. The second group approaches the process of predicting PPIs by combining several biological sources, usually in a classification framework. These two sets of methods are described below.

3.1.1 Methods Based on Single Biological Evidence

3.1.1.1 Gene Fusion

This method exploits the notion of gene fusion. Certain proteins, or domains (two separate proteins in a given species), may sometimes correspond to a single full-length protein in other species. This fused protein is called the Rosetta stone protein [80, 81]. Proteins that are fused in one genome are likely to interact in the other organism. The basis of this method involves searching for fusion events in a reference genome and inferring that the proteins that are fused in other genomes are either physically interacting or functionally related [80, 81]. Using this approach, Marcotte et al. [81] predicted PPIs in *E. coli*. They made two different predictions based on two ways of finding a gene fusion event. In the first method, they identified the gene fusion event by tracking the domain assignments of the proteins. Using this approach, 3,531 PPIs were predicted. In the second case, the fusion events are traced using sequence alignments; this method looks for whether two proteins in an organism can be aligned to a single protein in another organism. Based on this approach, the predicted set included 4,487 interactions. The analysis of the functions of the predicted pairs showed that the predicted pairs were closer to each other than the randomly paired proteins. Also 6.4% of their predictions were found to be known interacting proteins when compared to an experimentally identified PPI dataset. When there is a fusion event, this information can be very helpful; however, fusion events are rare, limiting the coverage of the method. Secondly, the ubiquitous domains, like the SH3 domains, may lead to false positive predictions as they are present in many proteins.

3.1.1.2 Gene Neighbor and Gene Cluster Methods

A set of approaches relies on the basic assumption that genes that interact or are functionally associated tend to be located in physical proximity to each other on the genome. This is thought to be due to the selective pressure to associate genes that are co-regulated. For instance, in prokaryotes, related genes are often co-localized into regions called 'operons'. Capitalizing on this information, this set of methods makes predictions based on the intergenic distances between genes, and/or based on gene orders [82–85]. Overbeek et al. applied this method to 24 bacterial genomes and found possible operon regions, then providing a list of functionally related proteins based on the predicted operons. Analysis by Huynen et al. [85] of the *Mycoplasma genitalium* genome showed that the fraction of genes that interact physically is 63% if conservation of co-regulation is required across six genomes. This number increases to 80% if the conservation of only three genomes is required. Dandekar et al. [82] showed that the fraction of genes known to interact physically was 75% in a set of conserved gene pairs in triplets of genomes that included at least two distantly related genomes. Similar results were obtained for yeast and worm [86, 87].

3.1.1.3 Phylogenetic Profile Methods

Phylogenetic profile methods also exploit the genomic context. These methods are based on the hypothesis that interacting proteins share a similar evolutionary history to preserve interactions and functionalities [88, 89]. A phylogenetic profile for each protein is constructed based on the presence or absence of that protein across a range of genomes. Genes that 'travel' together during evolution are assumed to be involved in similar cellular processes. Similarities of profiles are calculated, and those proteins with similar phylogenetic profiles are considered potential interacting partners [88–90]. Pellegrini et al. [88] applied this method to *E. coli*. All proteins are associated with 16 other genomes, and presence or absence of close homologues in these organisms is coded in a vector. These boolean vectors are then used to cluster *E.coli* genes. Their analysis shows that genes that participate in the same pathway or cellular component are likely to cluster together. The method can also be used for identification of domain-domain interactions. In this case a profile is constructed for each domain [91]. One obvious drawback of this methodology is that it fails to correctly classify ubiquitous proteins, i.e. proteins that are present in all genomes but are not necessarily functionally linked. Additionally, evolutionary processes such as gene duplication, loss, and horizontal gene transfer could hamper accurate construction of phylogenetic profiles.

3.1.1.4 Domain Profile Methods

A protein *domain* is part of a protein sequence representing an independent folded structure that can evolve, function, and exist with or without the rest of the protein. A *binding motif*, on the other hand, is a linear sequence motif that is recognized by a domain in one protein, being part of its binding site. These motifs are short sequence patterns with lengths of approximately ten residues that mediate binding to a common domain [92]. PPIs are often mediated through domain-domain interactions or domain-motif interactions. A number of methods have used this observation to infer PPIs [93–100]. The common theme in these methods is utilizing the statistics of the occurrence of domain-domain or domain-motif pairs in sets of interacting protein pairs and using these estimations to infer PPIs for new protein pairs.

The first approach that employed this idea was the association method [93], which scored domain pairs by their overrepresentation in interacting proteins of yeast. Later, Deng et al. [94] extended this model to all possible pairs of domains between a pair of proteins. They assumed that two proteins interact if and only if at least one pair of domains interacts from the two proteins. They also took into account that PPI data can be noisy. Their approach used a maximum likelihood estimation (MLE), where the probability of interaction for domain pairs is estimated by maximizing the likelihood of the observed PPI network. The above methods may preferentially identify promiscuous domain interactions, because they focus on those that occur with the highest frequency. Riley et al. [98] proposed the domain pair exclusion analysis (DPEA) method to extend the MLE approach. Their method assesses the contribution of each potential domain interaction to the likelihood of a set of observed PPIs from the incomplete interactions of multiple organisms. Iqbal et al. [100] addressed the problem of predicting protein domain interactions in yeast by using belief propagation. Belief propagation is a powerful message passing algorithm for probabilistic inference [101].

The major limitation of the domain profile method is its dependence on the accuracy and coverage of the domain assignment. Predictions can only be made for those proteins that have at least one domain assigned to them. However, many proteins have not been annotated with an identifiable domain.

3.1.1.5 Interolog Based Methods

Interolog approaches are based on transferring the knowledge of known interacting pairs across genomes to discover novel interactions. The rationale behind this approach is that if two proteins interact in one organism, their homologs in another organism have a higher chance of interacting. This is based on the assumption that sequence and structural similarities between gene products suggest functional similarities. A number of methods are based on mapping interologs onto other organisms through comparative genomics [97, 102, 103]. Yu et al. predicted interacting pairs in yeast *C. elegans*, *D. melanogaster*, *H. pylori*. They concluded that interlog predictions are feasible when the homology is larger than 80%.

3.1.1.6 Coevolution and Correlation of Phylogenetic Distances

This set of methods relies on the co-evolution of proteins at the sequence level. The approach relies on the accuracy of the observation that the interacting proteins must co-evolve to preserve their ability to interact with one another. Co-evolution of two proteins is quantified through the similarity of their phylogenetic trees [104, 105]. Authors observed that the phylogenetic trees for known interacting protein families tend to show a higher degree of similarity than non-interacting proteins [93]. Initially only for the two domains of phosphoglycerate kinase, Goh et al. [104] constructed trees based on distance matrices and then quantified the similarity of the trees based on the linear correlation of the distance matrices. Pazos et al. [105] extended this approach to larger sets of interacting proteins and protein domains. In their study, they initially identified sets of orthologus proteins in 14 genomes, using *E.coli* as the reference genome. In this analysis, they found that it is possible to predict the interactions of proteins based on strength of the correlation between the distance matrices of pairs of proteins. At a chosen correlation cutoff of the distance matrices, their model predicted 2,742 PPIs. The model's performance was evaluated by comparing it to that of a small set of experimental PPI known at the time.

3.1.1.7 Sequence-Structure Threading Based Methods

Another class of methods predicts interactions based on structural information. An example of such an approach was presented by Aloy et al. [106, 107], in which they derived statistical potentials from known interactions. Given a 3D complex and alignments of homologues of the interacting proteins, these statistical potentials were used to assess the fit of any possible interacting pair in the complex. These methods not only predict the presence or absence of PPIs, but also provide details of the interacting surfaces, such as identification of contacting residues. Lu et al. [108] applied the threading approach to the complete yeast genome. Each possible pairwise interaction among more than 6,000 encoded proteins was evaluated against a dimer database of 768 complex structures by using a confidence estimate of the fold assignment and the magnitude of the statistical interfacial potentials. They identified 7,321 pairwise interactions among 1,256 proteins. 374 of the 7,231 interaction were in agreement with the experimentally identified PPIs. Again, the quality of the predictions was estimated based on the cellular localizations and biological functions of the predicted interactors.

We should note that protein docking is also a commonly applied methodology; however, docking is not used to predict which proteins interact with each other but instead to characterize the physical details of the interactions. Protein docking relies on the search for the best geometrical and polar fit between the two interacting protein structures [109].

3.1.2 Methods Based on Multiple Sources of Biological Evidence

In order to make more accurate predictions, a second group of methods utilizes multiple types of evidences simultaneously. These information sources can include direct interaction information, such as noisy interaction data derived from high-throughput experimental results, as well as indirect information sources. For instance, interaction of two proteins with similar mRNA expression profiles is more likely; as discussed earlier, proteins with certain protein domain pairs are also more likely to interact [110]. The indirect information sources individually are usually weakly associated with the interaction but can yield reliable predictions when analyzed as a group. Studies that rely on multiple information sources typically formulate the problem as a binary classification task and solve the task with a classifier. In this framework, a classifier is trained to distinguish between positive examples of truly interacting protein pairs and negative examples of non-interacting pairs. Each protein pair is encoded as a feature vector, where features represent a particular information source regarding either protein interaction. Each of these methods intends to assess each features' predictive value on samples of known positives and negative examples. Thereafter, the model was extrapolated to genome scale, and the model predicts the chance of possible interactions for every protein pair using their associated features.

The advantage of the classification approach is that it allows the combination of highly dissimilar types of data (i.e., numerical and categorical) probabilistically; it can handle missing biological data, which is common in biological datasets, and it naturally assesses the importance of each information source according to its predictive power. Supervised binary classification requires a positive and a negative set of examples to learn the classifier function [146]. One challenge in defining negative example sets is their lack of a 'gold-standard' non-interacting proteins set. In some studies [111, 112] non-interacting protein sets are constructed based on cellular locations of the proteins; proteins that are localized in different parts of the cell are less likely to participate in an interaction. However, these approaches also are likely to yield their own biases [113]. A simpler approach of selecting negatives uniformly at random is therefore commonly preferred [96, 113–115].

For predicting direct PPIs or co-complex relationships, a number of statistical classifiers have been applied, including Naïve Bayes [112, 116], Bayesian networks [112], decision trees [114], kernel based methods [115], Random Forests [52, 68], and logistic regression [117, 118]. Below a few of these approaches and a study by Qi et al. [119] that compared different classifiers in a systematic fashion will be discussed.

Jansen et al. [111] applied the Bayesian network to predict complex memberships of proteins in yeast. Hence, their goal was to predict whether two proteins are in the same complex, not whether they necessarily had direct physical contact. Bayesian network is a graphical model approach, where nodes represent variables and directed edges between variables represent conditional probability relationships [101]. Features were derived from interaction data obtained from high-throughput experiments comprised of Y2H and *in vivo* pull down experiments, the correlation of mRNA amounts in two expression data sets, information about whether proteins are essential for survival, and annotations of the biological functions of genes. They derived the positive examples from the MIPS (Munich Information Center for Protein Sequences) complexes catalog [120], while the non-interacting pairs are synthesized from lists of proteins localized in different parts of the cell. Using these, they verified the model by comparing the predictions against held-out experimental interaction data, including the results of a TAP-MS study that became available at the time. The advantage of Bayesian networks is that they can be readily interpretable, as the structure of the Bayesian network represents dependency relations among information sources. They predict a protein pair as positive if its combined likelihood ratio exceeds a particular cutoff, and consider it as negative otherwise. The performance of the model is evaluated using a seven fold cross-validation protocol.

Rhodes et al. [116] predicted PPIs in human. They utilized protein domain assignments, gene expression measurements in human tissue samples, biological function annotations and orthologus PPIs. They derived a positive set of PPIs from the Human Protein Reference Database (HPRD) [121], a resource that contains known protein-protein interactions manually curated from relevant literature by expert biologists. The negative examples were generated based on sub-cellular locations, in which one protein was located in the plasma membrane and the other in the nuclear component. First, separated predictors were constructed based on each piece of information; next, a Naïve Bayes classifier was trained to combine the separate classifiers. The authors validated the accuracy of the predictive model on an independent test set of known interactions and experimentally confirmed two of the predicted interactions.

Qi et al. [119] provided a comparison between the multiple machine learning techniques in yeast PPI prediction. Importantly, authors also made a distinction between different prediction tasks: prediction of 1) physical interaction 2) co-complex relationship and (3) pathway co-membership. For each of these tasks, they compiled the appropriate positive examples from different databases. In these three separate tasks, six commonly used machine learning algorithms were compared: support vector machines (SVM), Bayesian networks, and decision trees, logistic regression, Random Forest, and k-nearest neighbor. Diverse features were encoded in two different ways. In the detailed encoding, every experiment was considered a feature whereas in the second encoding, similar experiments were grouped together and a single feature was used for each group. The authors concluded that in all three tasks, the Random Forest performed the best. Several factors could be the reason why the Random Forest classifier performs well in prediction of PPIs. The features are derived from biological sources, which are inherently noisy and features correlate strongly with each other. As the Random Forest classifier uses voted ensemble method and it includes several randomizations in selecting the training examples and in selecting the subsets of features to induce each tree in the forest (Section 5.2.2.2), it can prove more robust to missing features and noise. Finally, they found that the importance of different features depends on the specific prediction task and the way each feature is encoded. Correlation of gene expression was found consistently to be the most important feature for all three prediction tasks.

A recent study conducted by Mohammed et al. [122] focused on a different aspect of PPI prediction, that of minimizing labeling efforts via active learning techniques. As biological experiments are labor-extensive, obtaining true labels of data points is expensive. Active learning is a machine-learning method, in which the objective is to minimize labeling effort by judiciously selecting the examples to obtain labels. Initially, the classifier is provided with a few labeled instances and a large set of unlabeled instances. The active learner selects the most informative data point for the learning task from the unlabeled set, where informativeness is typically defined as maximal expected improvement in accuracy; the active learner then asks an oracle about this data point. The oracle is in this case the lab experiment. It returns the label whether the pair interacts or not. The new data are included into the labeled set and classifier is updated with the new training data. This process is repeated until a predefined budget, if any, is consumed or if the desired performance is achieved. In their setup, Mohammed et al. [122] tried four different active learning strategies for selecting the most informative data point for the task of predicting PPIs in human. Their results demonstrated that active learning enables better learning with less labeled training data. They have simulated the active learning setup; however, this strategy can be used in a real setup to couple computational and experimental efforts and minimize the efforts to obtain a good classifier.

3.2 Review of Prediction of Inter-species Protein-Protein Interactions

In contrast to computational methods applied to predict PPIs within a single organism ('intra-species prediction'), computational work on predicting PPIs between organisms ('inter-species prediction'), including between hosts and viruses, has been rare. The work presented in this dissertation was one of the first ventures in this research area. Only a subset of the methods applied in intra-species prediction described in the previous section is applicable to host-pathogen systems. For example, gene fusion and gene neighborhood methods have limited applicability, since there is no biological evidence that supports the assumption that two proteins in a pathogen are likely to interact if they exist as a single protein in a related pathogen. The phylogenic profile method might also be difficult to extend to host-pathogen systems since the pathogens also co-evolve with their hosts. On the other hand, techniques such as domain-profile methods or interolog-based methods have been applied to bacterial host-pathogen systems, and these are described below in detail. I believe the list of publications that have focused on predicting inter-species protein interactions provided below to be exhaustive.

3.2.1 Domain Profile Approach

Dyer et al. [123] proposed a method for predicting human-*Plasmodium falciparum* interactions. In their approach, they adapted the domain-profiles approach of [93] to predict host-pathogen PPIs. They first estimated domain-pair statistics from the human PPI network. From this set of human PPIs, they estimated the probability of two proteins' interaction given the domain pairs each protein is assigned. Using these estimated pairwise-domain statistics, they assessed how likely an interaction between pairs of human, *Plasmodium falciparum* proteins is. An interesting feature of this method is that they used the human PPIs to estimate these probabilities since more data are available in this case. Then, they transferred this knowledge to the task of predicting human, *Plasmodium falciparum*, where data are scarce. However, the general limitations of the domain interaction prediction methods discussed above are problematic here as well, the most severe limitation being that predictions can be only made for proteins assigned to domains. This especially limits the applicability of this method to viral proteins, as many viral proteins have few domain assignments.

Evans et al. [124] predicted HIV-1,human PPIs based on sequence motif-domain pairs, where the motifs are short eukaryotic linear motifs (ELMs) that mediate binding to a protein domain. This is one of the features that we have used in our supervised learning framework [125]. Using a similar strategy to that of our previous work, Evans et al. searched for ELM motifs in the HIV-1 viral protein sequence alignments and hypothesized that human proteins with the domain that binds to this motif are a likely to interact. Although ELM domain motifs are likely to capture the transient interactions between the viral and human proteins, it suffers severely from a lack of coverage. Both the ELM motifs and the domain assignments are poorly annotated. For example, only about 20% of the human proteome has domain assignments. In this study, the ELM-domain feature to be one of our least informative features, probably due to the problem of low coverage.

3.2.2 Interolog Based Approach

Davis et al. [126] presented a comparative modeling approach to predict PPIs for ten human-pathogen pairs; this extended their previous intra-species work [127]. The pathogens were all of bacterial origin, including mycobacteria, kinetoplastida and apicomplexa, which are responsible for 'neglected' diseases. In their protocol, for each pathogen protein-pair, template pairs were initially identified that were known to interact and to have their 3D structures solved as a complex. Next, homology models of the host-pathogen protein pair that bears similarity to these target pairs are used to build 3D structural models of the host-pathogen pair. Those with good scores are then filtered based on sub-cellular location and expression properties. The approach is limited by the coverage, since the pairs have to match a template protein pair.

Lee et al. [128] similarly applied an interolog-based prediction method to infer interactions between *Plasmodium falciparum* and human proteins. Having identified interolog-based predictions, they filter the initial list to match cellular localization constraints. For example, proteins that are in the nucleus both in human and *Plasmodium falciparum* are not likely to interact. These filters utilized protein location annotations and the presence of translocational sequence signal on protein sequence that is needed to translocate the protein to the red blood cell cytoplasm, where the host manipulation takes place.

Tyagi et al. [129] applied a similar homology search method. They focused on interactions occurring at the early stages of pathogenesis of *H. pylori*, that is the attachment of the pathogen to the host and the next immediate steps which require the recruitment of the secreted proteins of *H. pylori*. First the transmembrane human proteins are identified. Then homologs of the transmembrane proteins and the viral proteins are found by querying them using PSI-BLAST and RPS-BLAST against interaction databases. The pairs that achieve a cutoff of similarity were accepted as interacting. They predicted total of 623 *H. pylori* proteins with 6559 human proteins. The predicted interactions included 13 experimentally verified secreted proteins. In a later work [130], authors applied this method to predict PPIs between human and three pathogens *E. coli*, *Salmonella enterica typhimurium* and *Yersinia pestis*.

3.2.3 Structural Similarity Based Approach

Doolittle et al. [131] studied predicting HIV-1, host interactions based on structural similarities between HIV-1 and human proteins. Similar to the idea of mimicking human interaction partners in my work [125], their method was motivated by the structural similarities of the target proteins interaction partners and the HIV-1 proteins. For this purpose, the first pairwise structural similarities between host and pathogen proteins are retrieved. Those human proteins that contain regions with high structural similarity to an HIV-1 protein are referred to as 'HIV-similar'. Next, known interaction partners for these HIV-similar proteins are obtained. This list is referred to as 'targets'. This list is filtered by RNAi screens and cellular co-localization information. The advantage of this method is that it makes use of the structural information available. A limitation, however, is that for most of the HIV-1 proteins, only fragments of structures are solved.

Huang et al. [132] aimed to find the functional association network between Influenza A (H1N1) virus proteins and human proteins. Their assumption is that if the viral proteins and the human proteins were annotated with the same functional or molecular process annotations, they are likely to interact. When the authors constructed a putative prediction list based on this rule, the resulting network was large. They further filtered the network by looking at subsets of interactions. To identify these interesting subsets, they performed a k-core decomposition to analyze the core area of the network. K-cores are obtained by recursively removing all the vertices with degree smaller than *k*, until the degree of each remaining vertex is larger than or equal to k. The vertices with coreness equal to or greater than 4 were defined as core nodes. There are 101 core nodes, which include four virus proteins and 97 human proteins. The authors concluded that linkages between them formed the core functional association network. In order to assess the biological relevance of the core nodes, Huang et al. performed a gene ontology functional enrichment. However, as the network has been already constructed based on shared Gene Ontology terms, not surprisingly, the functional network was found to be overrepresented in certain functional terms.

Chapter 4

Thesis Overview

Identifying interaction partners of a protein often allows inferring its function in the cell. Similarly, knowing the set of protein interactions that take place between the virus and the host allows us to map the molecular details of the virus' manipulation of cellular processes and the host cell machineries that the virus depends on for a successful replication cycle. Such knowledge opens avenues for new therapeutic and preventative strategies. Therefore, detecting and characterizing PPIs between the host and virus have long been a focus of experimental biology. However, as reviewed in Section 2.2.1, there is no single, cost-effective, reliable experimental technique that allows high-throughput identification of PPIs. On the other hand, there is a large body of accumulated proteomic, genomic and phenotypic data on both the human cell and viruses, which could provide evidence on host-virus interactomes. If integrated properly; this body of data can accelerate the experimental efforts to identify PPIs.

Working towards defining host-virus interaction networks, this thesis aims to provide high quality curated interaction data, compile and identify biological information that serve as predictive features, predict novel host-virus direct PPIs and stratify the reported interactions. I focused on the HIV-1-human interactome because it is clinically important and represents the system with the richest experimental data available. However, the methods presented here can easily be extended to other host-virus systems as pertinent data become available. This chapter provides an overview of the thesis, summarizing open questions and challenges and the way in which they were approached.

4.1 Predicting Virus-Host Interactions

In this thesis, predicting PPIs was defined as a binary classification task, where each possible protein pair falls into one of two classes, the ’interacting protein pairs’ (positive class) and the ’non-interacting protein pairs’ (negative class). Predicting host-virus interactions requires in identifying biological information that can serve as predictive features. Several data types that have been useful in the intra-species prediction task (see Section 3.1) are not directly applicable to the host-virus setting; such data types include co-expression of genes, gene order and location. Therefore, identifying information that is predictive in distinguishing interacting protein pairs from non-interacting ones is important. The first challenge is that the biological information is scattered throughout different databases; furthermore, a large amount of information is not yet catalogued in databases. Through an extensive curation process, I identified experimental results pertinent to host-virus interactions and extracted the relevant data from databases or published articles. The first set of biological information assembled in this manner includes:

- Gene expression data of HIV-1 infected versus uninfected samples
- Gene ontology in terms of HIV-1 and human proteins annotated for the three different gene ontologies: biological function, molecular process and cellular location.
- Sequence information of HIV-1 and human proteins
- Datasets of motif and protein domains that mediate interactions among interacting protein pairs
- Posttranslational modifications of HIV-1 and human proteins
- Tissue expression of human proteins and HIV-1 susceptible cells and tissues
- Network properties of human proteins within the human interaction network. These include degree, clustering coefficient and network centrality of a vertex in the human protein-protein interaction graph.

35 different features were derived based on these datasets, which are described in detail in Chapter 5 and publication [125]. The model presented here was the first attempt

in literature to predict the global set of interactions between HIV-1 and human host cellular proteins. In encoding relevant biological information, the cellular context of the host cell is taken into account. For example, the similarity of the HIV-1 protein to the putative human proteins' interaction partner in terms of sequence, translational modifications, function, molecular process and cellular location are encoded. The results (Chapter 5) demonstrated that the features that take into account the cellular contexts of these human proteins are especially informative. For instance, network node properties of the human proteins in the human PPI network are among the most predictive features. Chapter 5 describes a random forest model trained and tested using this feature set. The learner is empirically evaluated and compared extensively to external experimental datasets. These datasets feature host proteins detected in budding virion and published genome-wide RNAi experiments that identified host factors that affect HIV-1's success of infection. 21 host proteins that were predicted to interact with one of the HIV-1 proteins were tested experimentally as to whether they colocalize with vpr and capsid via single live cell imaging techniques. The results of the colocalization experiments provided experimental support for many of the predictions.

4.2 Extended Model with New Feature Set

As new pieces of biological information became available during the course of this thesis, I incorporated them into the model together with the cellular context, such as known human protein complexes or human protein pathways (described in Chapter 6 and publication [133]). The new datasets included:

- Genome-wide RNAi screens in which the host factors required for infection are identified
- Affinity purification-mass spectrometry applied to HIV-1 proteins
- Sets of human proteins detected in budding virions
- Interactions of human proteins with other host viruses

These supervised models made use of a subset of HIV-1, human protein interactions deposited in the NIAID HIV-1, human protein interaction database [134, 135]. The database includes interactions curated from scientific literature, where interaction is described based on the keywords. However, the distinction between physical interaction, functional association or indirect interactions is not provided. In the first two computational models described in Chapters 5 and 6, I derived a subset for direct interactions, where I filtered for interaction pairs reported with certain keywords indicative of a direct interaction. In the second part of the thesis, I proposed a better solution for obtaining higher quality datasets; as explained below.

4.3 Refining Literature Curated Protein-Protein Interactions with Expert Opinions

Obtaining a negative set is even more problematic than defining a high-quality positive set of interactions. This is because one cannot conclude definitely that two proteins do not interact; the most that can be said is the proteins are not found to be interacting under the experimental conditions and the particular experimental method used. One common method employs annotations of cellular localizations when choosing negative examples for training and testing purposes. The protein pairs from different cellular locations are treated as negative examples, with the assumption that the cellular constraints most likely prevent the proteins from participating in a biologically relevant interaction [111, 112]. While this method leads to high quality negative sets of interactions, Ben-Hur et al. [113] showed that this method can result in biased estimates of prediction accuracy because the constraints placed on the distribution of the negative examples make the prediction task easier, leading to optimistic estimates of the accuracy. A second alternative is to create a negative set that has been created by pairing proteins uniformly at random from the set of protein pairs not known to interact [113]. This is rationalized by the fact that the probability that two randomly chosen proteins will interact is small and most methods are able to handle contamination from the small number of potential false negative matches. These positive and negative datasets constitute the labeled data for Chapters 5 and 6. Neither of these methodologies are ideal for creating high quality positive and negative datasets. In Chapter 7, I addressed this issue and obtained high quality positive and negative labels on HIV-1, host direct PPIs. For, I i) collected opinions of HIV-1 experts about the interactions reported in literature and ii) formulated a probabilistic framework to assign reliability scores to interactions based on the resulting noisy, subjective expert opinions.

Not all of the published interactions reported in the literature are equally well-supported by experimental evidence. Some interactions have been validated by multiple groups and techniques; other interactions have not been validated in this way. This is a general concern, Mackay et al. [136, 137] argue that many reports of PPIs are founded on 'insufficient data' generated by limited strategies; others challenge the assumption that literature-curated interactions are of high quality [136–142]. 44% of all the pairs in the NIAID database for HIV-1,human interactions are reported only in a single publication (see Chapter 7). The lack of follow-up studies - especially by labs other than the one that found the first evidence for interaction - hints at the possibility that for many of these interactions, there may not be sufficient experimental evidence to support their direct interaction. Assessing the data quality of PPIs from small-scale experiments requires a complex judgment about the methods and results of each specific study. Some experimental techniques more conclusively identify functional relations, while others more conclusively identify direct interactions; techniques do not work uniformly well across all proteins (see Section 2.2.1). In addition to the variability in the powers and limitations of each technique, the condition under which a study is conducted, such as *in vitro* or *in vivo* environment, the strains used, the mutations introduced, if there are any labels introduced and if yes the attachment sites of the labels all represent potentially important factors. Such parameters should be taken into account when interpreting the results. Such a complex judgment can only be provided by domain experts. In order to arrive at reliability scores for the HIV-1, human PPIs, I took a crowd-sourcing approach. HIV-1 experts were presented with the accumulated published evidence and asked to annotate interacting pairs with labels based on whether they think the interaction is supported with enough evidence to conclude that the pair represents a direct physical PPI.

In cases where an interaction received multiple opinions from different experts, disagreements among the experts were common. This is true especially when there is not enough evidence accumulated to give a conclusive answer. Additionally, disagreements among experts might arise because of their biases, expertise and/or stringency levels;

e.g., some experts are more difficult to convince with partial evidence or with the results of certain experimental techniques. For these reasons, expert opinions are noisy and subjective. I thus formulated this as a computational problem: given noisy opinions and with varying numbers of judgments for each protein pair, how to accurately decide which of the expert-annotated pairs are more likely to have 'direct physical interactions' and the degree of uncertainty of those conclusions in the absence of a ground truth for the labels. I took a maximum likelihood approach to estimate the experts' labeling accuracies for each label type. Next, these estimated labeler accuracies were used to calculate the probability that the interaction is a true direct interaction. In this model, I did not assume annotators to have the same labeling quality; moreover, I took into account that experts may have different labeling qualities for the label types 'interacting' and 'non-interacting'. The computational model is provided in Chapter 7. It is not limited to curated data for HIV-1 protein interactions, but is applicable to other cases where multiple noisy labels needs to be combined, which is a common setting of crowdsourcing applications. The results showed that negative data obtained in this way especially improves model quality.

4.4 Multi-Task Learning for Virus, Host PPI Prediction

In the classifiers built to predict HIV-1,human PPIs, I pooled all the viral protein's interaction data together and solved the problem as a single task. However, the viral proteins undertake different functions and participate in different parts of the replication cycle, which implies they might be drawn from different distributions. This necessitates building different models for each viral protein. However, the lack of sufficient data for many of the HIV-1 proteins impedes the construction of separate models for each task. In order to overcome the data scarcity issue while not disregarding possible differences in data distribution across viral proteins, a multi-tasking learning strategy was developed. In this model, single tasks (learning the protein-protein interactions of each viral protein) were grouped based on their relatedness, where relatedness was based on their functions in the viral replication cycle [143]. The model is modified in the training phase as follows. In the random forest classifier, the training examples are bootstrapped when building the decision trees. During the bootstrapping step, the training examples are drawn from a modified distribution where the probability of each example being drawn is proportional to its relatedness to the viral protein at hand. Such a multi-task framework leads to more accurate predictions compared to single tasks, where only the tasks' training examples are used and the pooled task where all the training examples are used. The focus of this thesis was predicting virus-host phyical PPIs. To achieve this we provided computational methods and high-quality data sets to serve as features and labeled data. Applying our methods on HIV-1, virus system resulted with experimentally testable hypotheses on putative host-virus PPIs, some of which have been already validated experimentally.

Chapter 5

Predicting the HIV-1,Human Protein Interactome

5.1 Overview

Human immunodeficiency virus-1 (HIV-1) in acquired immune deficiency syndrome (AIDS) relies on human host cell proteins in virtually every aspect of its life cycle [69]. Experimental efforts have identified set of host factors that assist HIV-1 during the different steps of its replication cycle [144]. Nevertheless, the complete physical interactome between the viral and the human cell proteins is still far from complete. The model presented here was the first attempt to predict the global set of interactions between HIV-1 and human host cellular proteins [125]. I adopted a supervised learning framework, where multiple information data sources were utilized, including co-occurrence of functional motifs and their interaction domains and protein classes, gene ontology annotations, posttranslational modifications, tissue and gene expression profiles, topological properties of the human protein in the interaction network and the similarity of HIV-1 proteins to human proteins' known binding partners. A Random Forest classifier with this extensive feature set was trained and tested . The model's predictions achieved an average Mean Average Precision (MAP) score of 23%. The rank-ordered lists of predicted interacting pairs are a rich source for generating biological hypotheses and many of the predictions were experimentally validated.

5.2 Methods

5.2.1 Problem Setting and Formulation

Predicting physical interactions between HIV-1 and human protein pairs is formulated as a binary classification task and solved using supervised learning techniques. Specifically, data points are pairs of HIV-1 and human proteins indexed by $i = 1, \ldots, N$, where N is the number of all possible pairs between HIV-1 and human, which is 353,778 (number of HIV-1 proteins[1] × the number of human genes). The set of all possible pairs is S, which can be viewed as the set of edges in a bipartite graph, whose vertices on one side are the HIV-1 proteins and on the other side the human proteins. We refer to the set of known interaction pairs as $L^+ \subset S$. On the other hand those pairs whose label are not known will be referred to as $U \subset S$ and $U = S \setminus L^(+)$. A negative set is generated from the unlabelled set, the set of negative examples are L^-, where $L^- \subset U$. We referred to the union of the positive and negative examples as $L = L^+ \cup L^-$. $\forall$ protein pairs $i \in S$ a d-dimensional feature vector, $\mathbf{x^i}$, is constructed. Each of the d features is derived from one or more biological information sources. Each $\mathbf{x^i}$ maps to one of the two class labels, $y = \{$'interacting','non-interacting'$\}$. For those pairs in L, we know their class labels, Y_L. Given the feature matrix for L, X_L, and class labels, Y_L, we seek to learn a classifier $h : \mathcal{X_L} \rightarrow Y_L$ that will correctly predict the class labels of unseen data.

5.2.2 Classification

A Random Forest classifier was employed [145] to solve the binary classification problem. The Random Forest method was chosen based on its robustness in scenarios, where the features are noisy and redundant as is the case for the virus-host PPI prediction task. Furthermore, previously it has been demonstrated to outperform other well known supervised techniques in predicting intra-species PPIs [117, 119]. The Random Forest is

[1] gag p1 and gag p2 are excluded from the model due to the limited number of interactions and information available for these proteins.

an ensemble learning method, where multiple decision tree learners are bagged. Below, we first provide a brief description of the decision tree classifier, next we provide details of the Random Forest classifier.

5.2.2.1 Decision Tree Classifier

The decision tree classifier is a supervised learning technique which uses a sequence of decisions [146]. A decision tree is formed by a root node, a set of interior nodes and terminal nodes, which are also named as the *leaf nodes*. The root node and the interior nodes are collectively referred as *non terminal nodes*. Each non terminal node in the tree represents a test on an input feature and each descendant node divides the feature space into sub spaces based on the possible answers to this feature-value test. An instance is classified by a set of rules that is determined by the path starting from the root node, moving down the tree and ending in a leaf node. The arrived leaf node denotes the final classification for that instance.

Even for a small number of nodes in a tree, learning the optimal structure that will minimize a loss function is usually computationally infeasible due to the combinatorially large number of solutions. Most algorithms use greedy optimization by constructing the trees top down beginning with selecting the feature that classifies the examples best. In deciding the best splitting feature in a decision tree, impurity measures are commonly used [147, 148]. At each step a new feature and threshold is picked and based on the chosen feature-value test, the feature space is divided into the subspaces, represented by two new descendant nodes. This is recursed at each subtree until the stopping criterion is met e.g. reaching a preset minimum number of examples at a node. Decision trees with sufficient depth usually exhibit low bias and can capture complex feature interactions in the data [149].

5.2.2.2 Random Forest Classifier

In contrast to the decision tree learner, which is formed by only one tree, the Random Forest classifier bags several trees [145]. The motivation is to average many noisy but approximately unbiased models and in this way reduce the variance [149]. Let B be the number of trees in the forest and let the training data contain N examples and d features. To construct each tree in the bag, $b \in 1..B$, first a bootstrap sample, Z_b^*, of size N is drawn from the training set with replacement. Next, a tree, T_b, is learned using this bootstrap sample. In contrast to regular decision trees, where the best splitting feature is selected among all d features, in a Random Forest tree the best splitting feature is selected from a random subset of $m \leq d$ features. To classify a new example, the instance is sorted down on each tree and each tree gives a vote on what the predicted class label should be. The Random Forest classifies the new example based on the majority vote of the trees in the forest. Let x be the feature matrix for a new example, the label for a new example x will be:

$$\hat{y}^B = \text{majority vote}\{(\hat{y}_b(x)\}^B \tag{5.1}$$

where $\hat{y}_b$ be the class prediction of the b^{th} tree in the Random Forest. The Random Forest algorithm is detailed in Algorithm 1.

Algorithm 1 Random Forest algorithm for classification [145].[2]

```
Let $N$ be the number of examples in the training data, $d$ the number of features, $B$ the
number of trees in the forest, $n_{min}$ minimum number of examples allowed on a node,
$m$ number of features to be used for determining the splitting feature.
1. Construct Random Forest:
for $b = 1$ to $B$ do
    2. Construct a Random Forest tree $T_b$:
    a) Draw a bootstrap sample of $Z_b^*$ of size $N$ from the training data.
    b) Grow a tree $T_b$ using $Z_b^*$. In growing the tree at each terminal node of the tree
    recursively apply the following steps:
    repeat
        i. Select $m \leq d$ features among the $d$ features at random as candidates for splitting.
        ii. Pick the best splitting feature among the $m$ features based on Gini impurity
        index.
        iii. Split the node based on the chosen feature into two.
    until $n_{min}$ is reached
end for
2. Random Forest is the ensemble of trees $\{T_b{}^B\}$
3. Let $\hat{y}_b$ be the class prediction of the $b^{\text{th}}$ tree in the Random Forest, then the label for
a new example $x$ will be:
$\hat{y}^B = \text{majority vote}\{(\hat{y}_b(x)\}^B$
```

The *impurity* of a node captures how dissimilar the instances on this node are to each other in terms of class labels. In constructing the trees in the Random Forest, the Gini impurity index is used as the splitting criterion [145]. The splitting feature is chosen as the one that reduces the Gini impurity index the most. Specifically, the *Gini impurity index* associated with node t for feature x_j, which takes r values $\{q_1 \ldots q_r\}$:

$$\text{Gini}(t, x_j) = \frac{n_1}{N_t} I_{gini}(t, x_j(q_1)) + \frac{n_2}{N_t} I_{gini}(t, x_j(q_2)) + \ldots + \frac{n_r}{N_t} I_{gini}(t, x_j(q_r)) \tag{5.2}$$

where node t has N_t examples and is split into r descendants, n_i is the number of examples at the descendant node i [150]. $I_{gini}(t, x_j(q_i))$ is the *Gini index* for the node t, associated with feature x_j when it takes a value of q_i and is computed as:

$$I_{gini}(t, x_j(q_i)) = 1 - \sum_{c \in Y} p_c(x_j(q_i))^2 \tag{5.3}$$

where $p_c(x_j, q_i)$ denotes the proportion of examples whose x_j feature takes the value q_i and belongs to class c. Y denotes the set of all possible class labels. The splitting criterion is based on choosing the attribute with the lowest Gini impurity index of the split. The Gini impurity index attains its minimum if all instances at a node belong to only one class. In that case, the node is pure and the misclassification rate is zero. On the other hand, it is at its maximum if each class has equal frequencies.

In our experiments, the Berkeley Random Forest package implementation was used [145], 200 trees were bagged. To cope with the unbalanced class distribution of examples, the cost of misclassifying a positive ('interacting') example is weighted more by a factor of w as compared to misclassifying a negative ('non-interacting') example. The model parameters, m - number of feature candidates at each node (see Section 5.2.2.2), w - the relative error cost of negative class to positive class - and parameters specific to feature encodings were tuned via 3-fold cross-validation on the training data. The parameters and features maximizing the mean average precision score were chosen.

Group 1 keywords: acetylated by, acetylates, binds, cleaved by, cleaves, degraded by, degrades, dephosphorylates, interacts with, methylated by, myristoylated by, phosphorylated by, phosphorylates, ubiquitinated by

Group 2 keywords: activated by, activates, antagonized by, antagonizes, associates with , cleavage induced by, causes accumulation of, co-localizes with, competes with, complexes with, cooperates with, decreases phosphorylation of, deglycosylates, depolymerizes, displaces, disrupts, downregulated by, downregulates, enhanced by, enhances, enhances phosphorylation of, enhances polymerization of, enhances release of, excludes, exported by, facilitated by, fractionates with, glycosylated by, imported by, inactivates, incorporates, induces, induces acetylation of, induces accumulation of, induces cleavage of, induces complex with, induces phosphorylation of, induces rearrangement of, induces release of, influenced by, inhibited by, inhibits, inhibits acetylation of, inhibits induction of, inhibits release of, inhibits release of, isomerized by, mediated by, modified by, modulated by, modulates, palmitoylated by, processed by, polarizes, promotes binding to, protects, recruited by, recruits, redistributes, regulated by, regulates, regulates import of, relocalized by, relocalizes, requires, sensitizes, sequesters, stabilizes, stimulated by, stimulates, synergizes with, transported by, upregulated by, upregulates

Table 5.1: List of Group 1 and Group 2 keywords. Group 1 keywords are those that most likely represent direct physical interactions and Group 2 set contains all the other keywords in the database. Interactions reported with Group 1 keywords are considered as direct PPI set.

5.2.3 Dataset

Positive examples (interaction class): Protein interactions between HIV-1 and human proteins reported in the literature were cataloged and are available in the NIAID HIV-1,Human protein interaction database (NIAID database) [134, 135]. These interactions were manually extracted from publications by reviewing more than 100,000 articles. This interaction network includes 1448 human proteins that interact with HIV-1 proteins comprising 2589 unique HIV-1,human PPIs. Table 5.2 lists the number of interactions per HIV-1 protein. Since our aim is to predict binary direct physical interactions, we need a set of direct interactions. However, such information is not listed in the NIAID database. Instead, the database describes each interaction by one or more descriptive key phrases, which are extracted from publications reporting these interactions. Some of these keywords are more likely associated with direct physical PPIs than others (e.g. 'interacts with' as compared to 'causes accumulation of'). We grouped the keywords into two exclusive sets: Group 1 keywords are those that most likely represent direct physical interactions such as 'interacts with' or 'binds' and Group 2 keywords which contain all

HIV-1 protein	**Number of HIV-1-Human Protein Interactions**	
	Group 1 type	Group 2 type
Env gp41	37	118
Env gp120	195	336
Env gp160	54	121
Gag capsid	19	13
Gag matrix	39	37
Gag nucleocapsid	5	19
Gag p6	14	0
Gag pr55	15	32
Nef	71	119
Integrase	72	6
Protease	60	18
Reverse transcriptase	17	22
Rev	33	29
Tat	336	420
Vif	54	10
Vpr	35	134
Vpu	7	13
Total	1063	1454
Number of unique human proteins involved	721	914

Table 5.2: The number of interactions between HIV-1 proteins and human proteins according to the NIAID HIV- 1, Human Interaction database [134, 135]. An interaction is classified as a Group 1 type of interaction if it is described by at least one of the Group 1 keywords and classified as Group 2, if otherwise.

the other keywords in the database. The list of Group 1 keywords are listed in Table 5.1. An interaction is defined as Group 2 type of interaction if it is not described by any Group 1 keyword. The Group 1 interactions constituted the 'interaction class', L^+. Group 2 was not used during the learning phase. Group 1 interactions includes 1063 pairs involving 721 human proteins, whereas Group 2 interactions were 1454 involving 914 proteins. Figure 5.1 displays the network representation of the two groups of interactions.

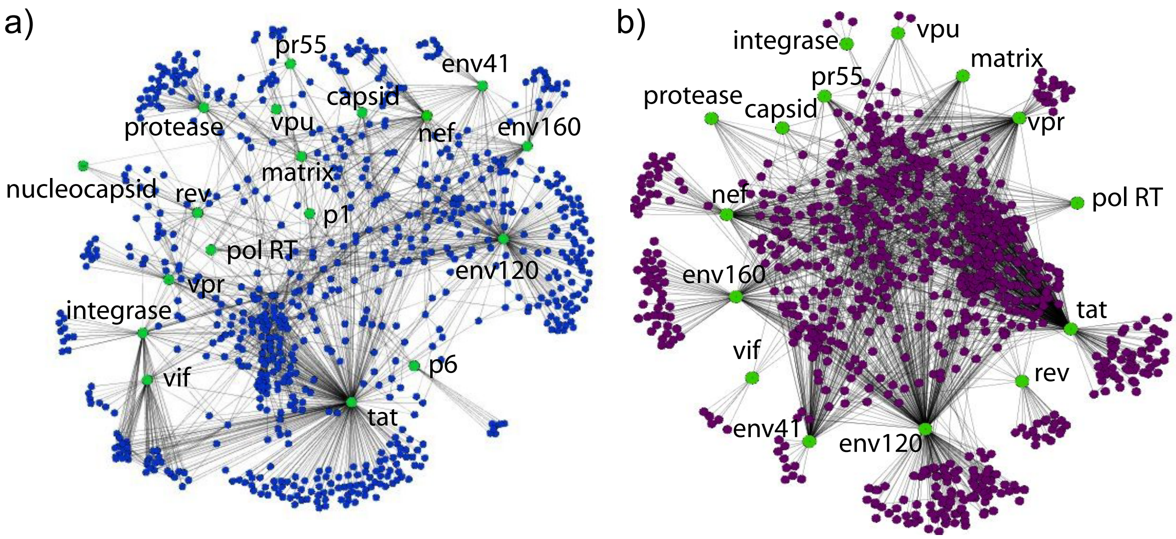


Figure 5.1: The HIV-1 human interactome defined by a) Group 1 interactions b) Group 2 interactions. HIV-1 proteins are colored green, human protein in Group 1 and Group 2 interactions are colored blue and purple respectively. The abbreviated protein names are as follows: env120:envelope protein env120, env160: envelope protein env160; env41: envelope protein gp41; p1: gag p1; p6:gag p6; pol RT: pol reverse transcriptase. The network visualizations are created by Cytoscape software [151].

Negative examples (non-interaction class): Since it cannot be proven that two proteins do not interact, there are no negative sets available for PPI prediction tasks in general. For training and testing purposes, a common method to create such a negative dataset is to choose protein pairs uniformly at random from the set of protein pairs which are not known to interact and treat them as non-interacting protein pairs [113]. This is rationalized by the fact that the probability that two randomly chosen proteins interact is small and most methods are able to handle contamination from the small number of potential false negative matches. We applied this strategy and selected negative example set, L^-, by randomly pairing human and HIV-1 proteins, after removing positive interactions from the universal set of all possible interactions (from the set $S \setminus L^+$). Together these two datasets constitute the labeled data $L = L^+ \cup L^-$. The ratio of the negative to positive class is assumed to be 100:1, a value chosen based on the average number of interactions involving HIV-1 proteins. In training, the negative to positive example ratio is treated as a parameter that is optimized through cross-validation.

5.2.4 Features

Source	Data obtained	URL
GO [152]	Gene ontology to describe genes' functions, processes and cellular locations	www.geneontology.org
EBI [153]	Human genes GO annotations are obtained	www.ebi.ac.uk
ELM [154]	Eukaryotic linear motifs, which mediate binding to a domain or a protein class	http://elm.eu.org
InterPro[155]	Domain assignments for human proteins	www.ebi.ac.uk/interpro
GEO [156]	Gene expression profiles for HIV-1 infected vs uninfected samples	www.ncbi.nlm.nih.gov/geo
HUPA [157]	Tissue expression and post-translational modifications of human proteins	www.humanproteinpedia.org
HPRD [121]	Post-translational modifications of human proteins and interaction data used	www.hprd.org
UniProt [158]	Human protein sequences and GO annotations for HIV-1	www.uniprot.org
Los-Alamos HIV Database [159]	HIV-1 protein sequences, alignments	www.hiv.lanl.gov
dbPTM [160]	Post translational modifications of human proteins	http://dbptm.mbc.nctu.edu.tw

Table 5.3: Biological data sources used in deriving features.

In order to extract features that can be informative to discriminate the two classes, a total of 35 features was extracted from various biological sources. The biological information is scattered in a wide range of data sources, which requires extensive curation efforts. A list of these biological sources is given in Table 5.3.

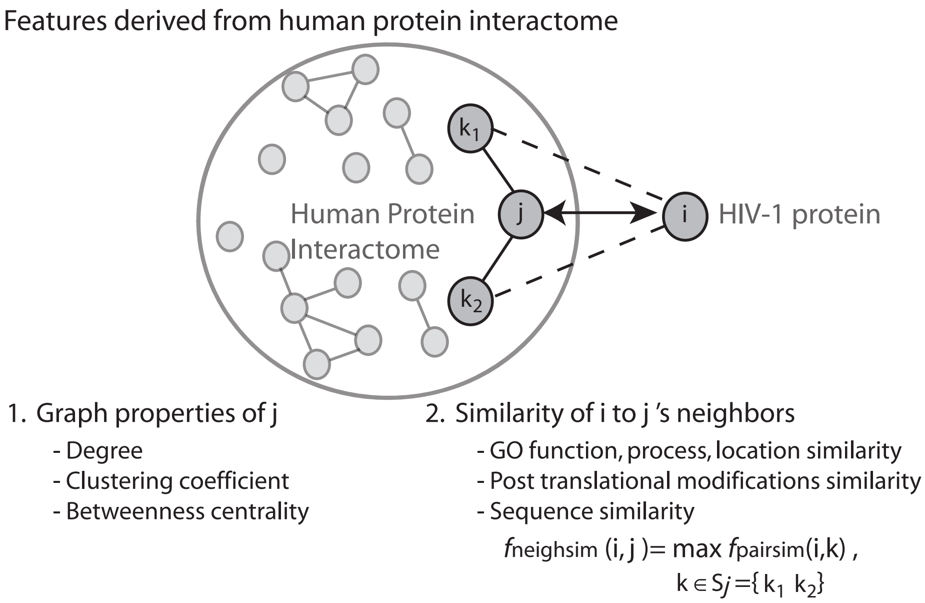

Features derived from human protein interactome
Human Protein Interactome
k1
j
k2
i
HIV-1 protein
1. Graph properties of j
- Degree
- Clustering coefficient
- Betweenness centrality
2. Similarity of i to j 's neighbors
- GO function, process, location similarity
- Post translational modifications similarity
- Sequence similarity
fneighsim (i, j)= max fpairsim(i,k) ,
k ∈ Sj ={ k1 k2}

Figure 5.2: Schematic showing features that incorporate knowledge of the human protein interactome. These features include: 1) graph properties of human protein j in human protein interaction network, which include degree, clustering coefficient and betweenness centrality of node j 2) the similarity of the HIV-1 protein, i, to human protein j's interaction partners denoted by $f_{\text{neigh}}(i, j)$ in the figure. In calculating the neighbor similiraity, the maximal similarity among the neighbors is used. Five features are derived this way; GO function, process and location similarity in addition to post-translational modification and sequence similarity.

A subset of features are specific to the HIV-1, human protein pair, whereas some encode information only about the human protein or only about the HIV-1 protein. The former has the potential to give information about whether the specific HIV-1, human protein pair interacts or not. On the other hand, non-pair specific features encode whether the human protein interacts with any of the HIV-1 proteins, or whether an HIV-1 protein interacts with any of the human proteins. Among the features two interesting (overlapping) groups of features are those features that make use of the human PPI network and features that encode the HIV-1 proteins' similarity to its human interaction partner:

Features with human interactome knowledge: The proteins the pathogen will target should in principle depend on interaction relationships between human proteins because the virus makes use of the existing communication pathways within the cell. Therefore, a set of features are used to encode the relationships of the host proteins within themselves.

Human interaction partner similarity features: The HIV-1 proteins that will interact with a human protein h, may bear resemblance to human protein physical interaction partners in the human proteome. To capture this property, these features encoded the HIV-1 proteins' mimicry of the human protein's interaction partners as illustrated in the schematic Figure 5.2.

Below, we detail the features used and the rationale behind them. Table 5.4 summarizes the features encoded, together with the groups they belong to. For all features, where appropriate, the missing values were substituted with the mean or median (for non-categorical and categorical attributes, respectively) of the available values for that feature.

5.2.4.1 Gene Ontology Features

The Gene Ontology (GO) [152] provides a defined vocabulary of protein attributes for molecular function, cellular component and biological process. In GO, each ontology is represented by a hierarchical directed acyclic graph (DAG), in which nodes are GO terms and edges are relationships between these terms. A term in the root is more general; the lower in the tree, the more specific the term. A child term may be an 'instance' of its parents' term ('is-a' relationship) or a component ('part-of' relationship). Genes are annotated with one or more terms. For each of the three ontologies we developed two features: 'pairwise GO similarity' measures the similarity between the HIV-1 and human proteins in a pair, while 'neighbor GO similarity' refers to the similarity between the HIV-1 proteins and the human protein's human interactors. To calculate the similarity between two protein annotation sets of GO terms, we employed the G-SESAME method [161]. This method compares the subgraphs of GO terms (starting from the specific GO term ending in a root term). G-SESAME not only considers the common ancestors the GO term pair have but also the location (closeness to the most specific term) and the relation type of the edges. Then, the GO semantic similarity between two proteins' annotation set is calculated by averaging the maximal similarity of each term to the other protein's annotation set.

Feature(s) name	Num. features	Feature type	Coverage
GO pairwise function similarity	1	HV	65.4
GO pairwise process similarity	1	HV	63.3
GO pairwise component similarity	1	HV	66.7
GO neighbor function similarity	1	HV, HPPI	42.6
GO neighbor process similarity	1	HV, HPPI	45.3
GO neighbor component similarity	1	HV, HPPI	45.3
Post translational modification similarity	1	HV, HPPI	40.0
Degree	1	H,HPPI	45.3*
Clustering coefficient	1	H, HPPI	20.0*
Betweenness centrality	1	H, HPPI	31.6*
Neighbor sequence similarity	1	H, HPPI	45.3
Pairwise sequence similarity	1	H	100.0
ELM, ligand feature	1	HV	2.3*
Gene expression features	4	H	44.0
Tissue expression	1	H	66.6
HIV protein type features	17	V	100.0

Table 5.4: Feature set derived for prediction of interactions between HIV-1 and human proteins. The first column lists the name of feature group, the second column gives the number of features of this group. The third column, describes whether the feature is specific to the HIV-1 protein pair (HV), only to the human protein (H) or only to the HIV-1 protein (V); and features that makes use of the human protein interaction network knowledge are also indicated in this column by 'HPPI'. The fourth column presents the percentage of pairs for which information is present (coverage). For gene expression features, the average coverage across the four gene expression data sets is given. In some of the features, coverage is 100 % as a result of the way the feature was encoded. For example, for the ELM-ligand feature, if the condition for the pair is not satisfied, the feature value for that pair takes a value of zero. In such cases, the percentage of non-zero elements is given. The features for which this applies are marked with * in the last column.

5.2.4.2 Graph Topological Properties

Three features measure the topological properties of the human protein in the human PPI network. The human PPI network is described as an undirected graph: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a set of vertices $\mathcal{V}$ (proteins) and a set of edges $\mathcal{E}$ (interactions). We utilized three topological properties of a vertex (human protein):

- **Degree**, k_v, of a vertex v is the number interactions (edges) in which it participates.
- **Clustering coefficient**, C_v, of a vertex(protein) v, measures the extent to which the protein's interaction partners are connected to each other and is defined as $C_v = 2n_v/k_v(k_v - 1)$. In this equation, k_v is the number of interaction partners (degree) of v and n_v is the number of edges present among its neighbors. Clustering coefficient is defined for vertices with degree $k_v \leq 2$.
- **Betweenness centrality** is defined as in reference [162]; for a node it is calculated as the fraction of shortest paths between node pairs that pass through the node of interest. High betweenness centrality indicates that the protein has control over the information flow between other proteins in the network.

5.2.4.3 ELM-Ligand Feature

One way of achieving PPI is through domain binding linear motifs, where a globular domain in one protein recognizes a linear peptide from another, creating a small contact interface. These motifs are short sequence patterns around ten residues long which mediate binding to a common domain (see Section 3.1.1.4). For example, the sequence pattern PXXDY (ELM id: LIG_SH3_5) is recognized by SH3 domains. Functional sequence motifs were downloaded from the Eukaryotic Linear Motif (ELM) database [154]. Based on the motif descriptions, we identified 109 ELMs, which mediate binding to a protein domain or a specific protein class. The feature evaluates whether an ELM motif is found in a given HIV-1 sequence and its ligand domain is present in the human protein. The HIV-1 sequences mutate rapidly and some motifs are very short. To avoid false positive matches, we only considered a motif match if it is conserved in multiple HIV-1 sequences

[125]; the feature value was weighted with the specificity of the motif, the weight was chosen through cross-validation.

5.2.4.4 Gene Expression Features

GDS ID	Description	Number of Features Derived	Reference
GDS2649	20 CD4+ and 20 CD8+ samples from HIV-1 patients at different clinical stages and rates of disease progression.	2	[163]
GDS172	Samples from brain frontal cortex of HIV-seropositive patients with HIV encephalitis (HIVE). 28 samples: 16 disease and 12 control samples.	1	[164]
GDS171	Gene expression profile of proliferating normal peripheral blood mononuclear cells (PBMC) infected with HIV-1. 3 infected and 3 uninfected for each time samples for each time points: t=0,12,24,48,72 hours	1	[165]

Table 5.5: Gene expression data sets are derived from the Gene Expression Omnibus (GEO) of NCBI [156] available at http://www.ncbi.nlm.nih.gov/geo/. The dataset ids are given in the first column, which can be queried in GEO. The second column describes the dataset. In the third column, samples are compared. The fourth column lists the number of features derived from the respective datasets. The reference reporting the gene expression study is given in the last column.

Four features derived from three gene expression studies (see Table 5.5), which reflect differential expression patterns of human genes across HIV-1 infected vs. uninfected samples. The differential expression for a human gene i was calculated as the log fold change:

$$f_{diff}(i) = log \frac{\langle g_i^{(+)} \rangle}{\langle g_i^{(-)} \rangle}$$

In the above equation, $\langle g_i^{(+)} \rangle$ and $\langle g_i^{(-)} \rangle$ are the average expression of the gene, i, across infected and uninfected samples respectively. In the case of time series experiments, where expression is measured at different time points, samples at each time point are compared between the infected samples and uninfected samples. The time point where maximum absolute fold change occurs was used as the feature value. Thus, if the gene expression of the infected and uninfected samples were measured at n time points, $t_1, \ldots, t_n$, and $\langle g_{i,t}^{(+)} \rangle$ and $\langle g_{i,t}^{(-)} \rangle$ were the average gene expression value at time t for gene i, across infected and uninfected samples respectively, the feature for the time series is encodes as:

$$f_{diff}(i) = max_{t \in \{t_1, \ldots, t_n\}} log \frac{\langle g_{i,t}^{(+)} \rangle}{\langle g_{i,t}^{(-)} \rangle}$$

5.2.4.5 Tissue Expression Feature

This feature encoded whether the tissues that the human proteins are expressed in are susceptible to HIV-1 infection or not. The tissues which HIV-1 reported to infected were obtained from [166]. The tissues where each human protein are expressed were obtained from HUPA database [157]. The feature value was set to 1 if the human protein was expressed in one of the HIV-1 susceptible tissues.

5.2.4.6 Sequence Similarity Features

For each pair, two sequence similarity features were utilized: i) pairwise sequence similarity and ii) similarity to human proteins' human PPI partners. The motivation behind the sequence similarity feature was the fact that homo-oligomers including homodimers are frequent in protein structures. Similarly, two protein structures with similar sequences, likely representing similar structures, might interact as well. Pairwise sequence similarity was computed for each pair using the BLASTP package [167]. The best hit subsequence's $-\log(\text{e-value})$ was used as the similarity measure, where e-value was the expected value of the alignment score to be found merely by chance [167]. In calculating the similarity between the HIV-1 protein and the human protein's interaction partners, pairwise similarities were first calculated, then the maximum similarity score was chosen among these.

5.2.4.7 Posttranslational Modification Similarity to Neighbor

Some PPIs require the protein(s) to be in a certain posttranslationally modified state. For such cases, the HIV-1 protein will require to mimic the posttranslational modification (PTM) of the human protein's interaction partner. For example, the N-terminal myristoylation domain of Nef is highly conserved and plays an important role in interacting with calmodulin. One of calmodulin binding partners is the human protein NAP- 22/CAP23, which is also myristoylated [168]. The feature was encoded as a binary feature such that for a pair, the feature takes a value of 1 if at least one of the human protein's human protein binders shares at least one common PTM with the HIV-1 protein, otherwise -1. To this end, we collected the HIV-1 proteins' PTMs from the literature. Experimentally verified posttranslational modifications of viral proteins are curated from published scientific articles manually. The information on post translational modifications each human protein undergo was obtained from three different databases (see Table 5.3) and combined using controlled vocabulary of posttranslational modifications provided by the UniProt database [169].

5.2.4.8 HIV-1 Protein Type Features

To capture the frequency of interaction of each viral protein, HIV-1 protein type features(ptf) were used to indicate the HIV-1 protein identity. 17 binary features encoded which of the HIV-1 protein participates in the pair. That is for the feature encoding vpr's identity, all pairs that include vpr, this feature will be 1 and for other pairs it will be -1.

5.2.5 Performance Evaluation

Classifier performance was evaluated with 3-fold cross validation in 10 repeat runs. In each repeat, a different random negative set is used. At each cross-validation step one third of the examples are held out for testing. The training data was also further split into three, and 1/3 was used as the validation data. The parameters for each cross-validation step were optimized on this set. When evaluating the performance of a classifier on an imbalanced test set such as is the case here, computing accuracy is not useful because the majority class can easily be predicted by chance. Therefore, we evaluated the quality of our predictive model using two figures of merits which ignore the success on the TN rate: the receiver operating characteristic (ROC) curve and precision vs. recall curve [170]. The Mean Average Precision (MAP) score is used to summarize the precision vs. recall curve and the area under the ROC (AUC) scores to summarize the ROC curve [170]. As testing the whole list is a tedious task and experimental biologists prefer to proceed with the most confidently predicted pairs, typically the low FP region of the ROC curve is of particular interest in the PPI prediction task. To evaluate the model specifically in this region, the partial AUC scores AUC50, AUC100, AUC200 and AUC300 were determined, measuring the area under the ROC curve until reaching 50, 100, 200 and 300 FP predictions, respectively.

5.2.6 Assessing Features' Predictive Power

Gini feature importance was derived from the Gini index described in Section 5.2.2.2 and is the sum of all decreases in the forest due to a given feature, normalized by the number of trees in the forest.

5.2.7 GO Enrichment Analysis

GO enrichment of the human proteins involved in the predicted interactions was identified using Ontologizer 2.0 [171] using the child-term parent intersection method and using Bonferroni correction for multiple hypothesis testing.

5.3 Results and Discussion

5.3.1 Classifier Performance

We trained a Random Forest classifier with a rich feature set (Table 5.4) derived from several biological information sources (Table 5.3). The performance of the model was evaluated through 3-fold cross validation experiments (each cross validation experiment was repeated 10 times, in each experiment the negative interaction data is selected ran-

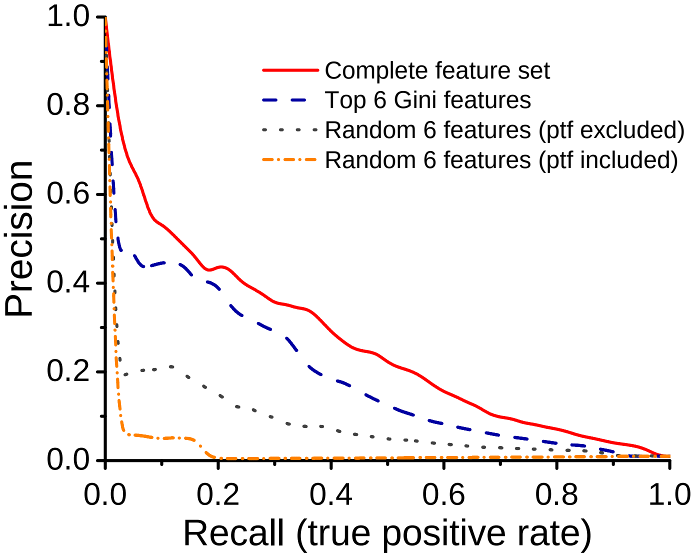


Figure 5.3: The average precision vs. recall curve of the Random Forest model trained on the complete feature set (solid red line), in comparison to models trained with a subset of features. The top 3 Gini features are degree, betweenness centrality, and GO neighbor process similarity features. The top 6 Gini features are the top 3 Gini features plus clustering coefficient, GO neighbor function, and location features. These are compared to two baseline classifiers, where 6 features were randomly selected from the set of features which does not include the top 6 Gini features, with and without protein type features (ptf).

domly). The average precision vs. recall curves of these experiments are given in Figure 5.3 (solid red line). Table 5.6 lists the average MAP, AUC and partial AUC scores of the model. The model achieves an average MAP score of $0.23(\pm 0.02)$. For PPI predictions, this is a very good performance, because of the highly skewed class distribution (ratio of positive:negative pairs of 1:100).

	MAP	**AUC**	**AUC50**	**AUC100**	**AUC200**	**AUC300**
Avg	0.2300	0.9150	0.0670	0.1073	0.1682	0.2156
Ste	0.0039	0.0022	0.0025	0.0030	0.0036	0.0042

Table 5.6: Averages (Avg) and standard error (Std) of MAP, AUC and partial AUC scores over 10 repeated 3-fold cross-validation experiments.

5.3.2 Features' Predictive Power

Biologically, it is of interest to identify the features contribute the most to the classification of protein pairs. This not only helps reveal relationships between different data sources, but can also suggest which data should be generated by experiments to find novel interactions in this and other host-pathogen systems. The feature importance was assessed based on the Gini importance of the Random Forest classifier (see Section 5.3.2). Strikingly, the graph property and the GO neighbor similarity features are ranked at the top, as shown in Figure 5.4.

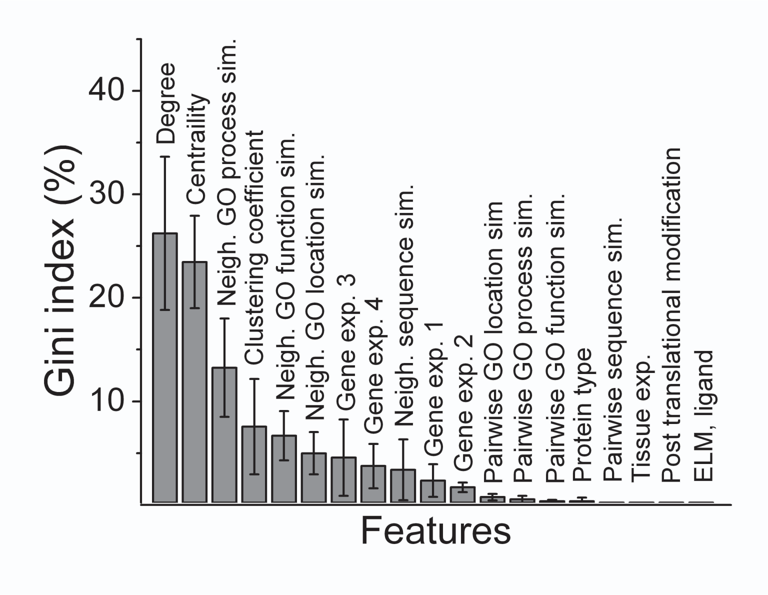


Figure 5.4: Gini importance indices for each feature. Protein type features are grouped together.

To assess the extent to which these features are predictive, we built models using the same train/test data splits as before with only the top 3 and top 6 Gini feature. The top 3 Gini features are degree, betweenness centrality and neighbor GO process similarity Figure 5.4 and the top 6 Gini features in addition include clustering coefficient, neighbor GO function, and cellular location similarities. These models were compared to two baseline models. In the first, the Random Forest classifiers were trained with 6 features selected randomly from the set of features excluding the top 6 Gini features. These random feature sets include the 17 protein type features (ptf), one for each HIV-1 protein. Since these vectors alone do not contain much information, this model forms a weak baseline. A second stronger baseline was built, where the 6 features are randomly selected from the set of features excluding ptf and the top 6 Gini features. Figure 5.3 compares the performance of the above 5 models. The top 6 Gini model performs quite strongly compared to both baselines. However, this model is not as good as the model built using the complete feature set. The top 3 Gini model performed significantly worse than the top 6 Gini model, but significantly better than the two baseline models suggesting that the additional top 3 features contain independent and complementary information. Statistical significance of these differences was confirmed by using on paired t-test comparison of the 30 experiments' MAP scores (at a significance level of 0.05).

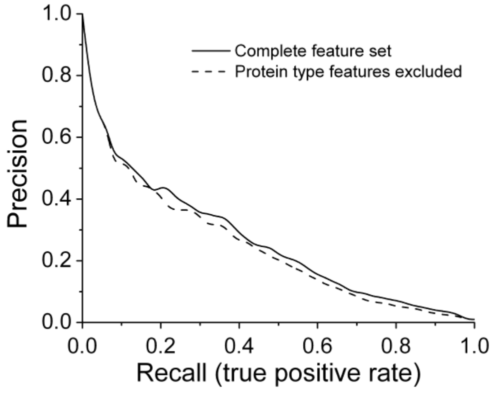


Figure 5.5: The precision vs. recall curves of the models trained with and without protein type features.

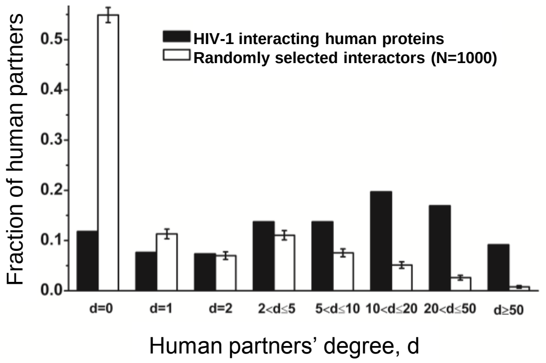


Figure 5.6: The number of human partners, degree, in the human PPI network (black) is plotted and compared with the degree distribution calculated over random graphs (white). The random graphs are generated by choosing each HIV-1 protein's neighbors uniformly at random from the all human protein set. 10.000 random graphs were generated

The above analysis reveals that the graph features and neighbor similarity features are very informative confirming our intuition in incorporating human interactome knowledge into the model. Graph properties have also been found previously useful in the intra-PPI network prediction task [20, 172, 173]. Furthermore, it has been proposed earlier that pathogens exploit network properties of the human interactome: it was shown that the Epstein-Barr virus targets high degree human proteins [41]. Similarly, Dyer et al. [174] reported pathogens tend to interact with host proteins with high degree and betweenness centrality. In analyzing the degree distribution of the HIV-1 proteins' human interaction partners, we also found an enrichment of hub proteins (see Figure 5.6). The significant performance difference between the top 6 Gini model and the complete model shown in Figure 5.3 indicates that the lower ranked features also contribute to the final performance. For example, the removal of ptf levels off the precision vs. recall curve with respect to the complete feature set (see Figure 5.5. The reason why some of these features' Gini importance scores are very low could be due to their low coverage (see Table 5.4).

5.3.3 Comparison with Other Biological Information

Precision	Recall	Total	Group 1	Group 2	Novel
0.51	0.20	3373	1045	243	2085
0.37	0.29	1948	1033	153	762
0.26	0.36	1442	1019	79	344
0.18	0.41	1087	889	41	157
0.13	0.47	630	543	18	69
0.09	0.47	284	247	9	28

Table 5.7: Number of predicted pairs at different choices of Random Forest score cutoff. Average recall and precision was calculated on the held-out test sets in cross-validation experiments.

A final model was trained with all available positive data. All HIV-1, human pairs were ranked according to their Random Forest score. The score measures the difference between positive and negative votes from the decision trees in the trained Random Forest model and reflects the margin of the decision. The derived ranke ordered list is available at URL www.cs.cmu.edu/~oznur/hiv/hivPPI.html. The set of predicted interactions depends on the chosen Random Forest score threshold; lowering the threshold will increase the TP rate at the expense of a higher FP rate. Table 5.7 presents the number of predicted interactions for different cutoff values (the precision recall values at that cutoff is given instead of the cutoff). At the lowest threshold considered (0), 2085 novel interactions are predicted, of which 1 in 5 interactions is expected to be true based on precision measured on the set of interactions withheld. Also reported in Table 5.7 group 2 predictions, which are those pairs that are reported in the NIAID database with weak keywords, but have not been used in model building. When also predicted as positive interactions, those associations reported in the literature are interesting predictions to pursue. The predictions referred to as 'Novel' in Table 5.7 are the predictions that are not reported in the NIAID database at all.

The predictions were also examined in light of the three siRNA screens that have been reported to have an effect on HIV-1 infection upon silencing [72–74]. Zhou et al. screen identified 291 human genes to be potentially important in the virus screen, whereas Brass et al. revealed 281 genes and König et al. revealed 295 genes. In addition to the overlap of the predicted human proteins, but I also examined whether the human protein is one of the reported siRNA genes' interaction partner in the human PPI network, since siRNA screens do not necessarily reveal direct interactions. We compared the predictions with a second type of biological data 'in virion'. These are the proteins detected in the HIV-1 virion [175]. There are 314 human proteins of them. Table 5.8 gives the size of the overlap of the model's predictions with these datasets. Although the comparison cannot provide means to verify the predictions; the overlapping pairs would be of particular interest to HIV-1 virologists: the siRNA data provide experimental evidence pointing at their functional relevance and the in virion overlapping set could help differentiate between mere by-stander human in virion proteins from those with functional roles for the virus. A subset of the predictions were followed up by colocalization studies.

Group 1

Precision	Recall	In Virion	Brass	Brass Interactor	König	König Interactor	Zhou	Zhou Interactor
0.51	0.20	153	36	351	63	110	37	383
0.37	0.29	150	36	351	63	110	37	382
0.26	0.36	147	36	347	63	109	36	376
0.18	0.41	135	34	331	60	106	33	360
0.13	0.47	82	18	244	42	89	21	274
0.09	0.47	35	8	142	23	51	9	159

Group 2

Precision	Recall	In Virion	Brass	Brass Interactor	König	König Interactor	Zhou	Zhou Interactor
0.51	0.20	36	4	125	7	55	12	154
0.37	0.29	26	3	87	5	40	6	104
0.26	0.36	13	1	46	3	21	2	53
0.18	0.41	6	1	27	3	12	1	29
0.13	0.47	4	1	14	2	6	1	15
0.09	0.47	2	0	8	0	3	0	8

Novel

Precision	Recall	In Virion	Brass	Brass Interactor	König	König Interactor	Zhou	Zhou Interactor
0.51	0.20	240	46	1054	77	422	73	1101
0.37	0.29	99	14	435	21	182	26	456
0.26	0.36	45	5	208	11	99	9	210
0.18	0.41	17	2	97	7	54	5	98
0.13	0.47	8	1	49	4	28	0	51
0.09	0.47	4	0	25	2	14	0	23

Table 5.8: Number of predicted pairs at different choices of Random Forest score cutoff. Average recall and precision was calculated on the held-out test sets in cross-validation experiments. The table presents the overlap (the number of the predicted pairs including the reported human gene) between the new predictions and three siRNA screens [72–74] and in Virion [175] datasets (for details, see text). 'Interactor' refers to the predicted interactions, where the human protein is at least one of the siRNA reported human protein's interaction partner. The comparison is divided into three groups Group 1, Group 2 and Novel.

5.3.4 Enriched Functions and Biological Processes

A global analysis of the predicted interactions by assessing the enrichment of GO functional terms in predicted Group 2 and novel interactions revealed 31 molecular processes, 19 biological functions and 14 cellular components ($p \leq 0.01$). Partial lists of significantly enriched GO terms for molecular process, function and cellular location are given in Tables 5.9, 5.10, 5.11 respectively. The full lists can be obtained in Supplementary Tables S5-S7 of [125]. For example, transcription regulator-, ligand-dependent nuclear receptor-, MHC class I receptor-, and protein kinase C activities are highly enriched molecular functions, while immune system process and response to stimulus are highly represented processes. Finally, macromolecular complex, membrane-enclosed lumen and plasma membrane are the top most significant cellular compartments.

GO term ID	GO term name	p-value
GO:0030528	transcription regulator activity	3.52e-38
GO:0005515	protein binding	2.97e-29
GO:0005488	binding	9.33e-20
GO:0008134	transcription factor binding	7.75e-18
GO:0003677	DNA binding	1.83e-13
GO:0004879	ligand-dependent nuclear receptor activity	1.85e-12
GO:0005057	receptor signaling protein activity	1.02e-08
GO:0016740	transferase activity	5.64e-08
GO:0032393	MHC class I receptor activity	5.83e-08
GO:0003676	nucleic acid binding	1.24e-07

Table 5.9: Enriched GO molecular function in the unique set of human proteins that involve novel and Group 2 predicted interactions.

GO term ID	GO term name	p-value
GO:0002376	immune system process	2.03e-54
GO:0051704	multi-organism process	2.11e-45
GO:0050896	response to stimulus	8.12e-44
GO:0043170	macromolecule metabolic process	3.83e-24
GO:0008150	biological-process	5.02e-24
GO:0019882	antigen processing and presentation	1.87e-19
GO:0032502	developmental process	6.36e-19
GO:0048518	positive regulation of biological process	1.73e-17
GO:0044419	interspecies interaction between organisms	2.47e-16
GO:0010467	gene expression	4.30e-16

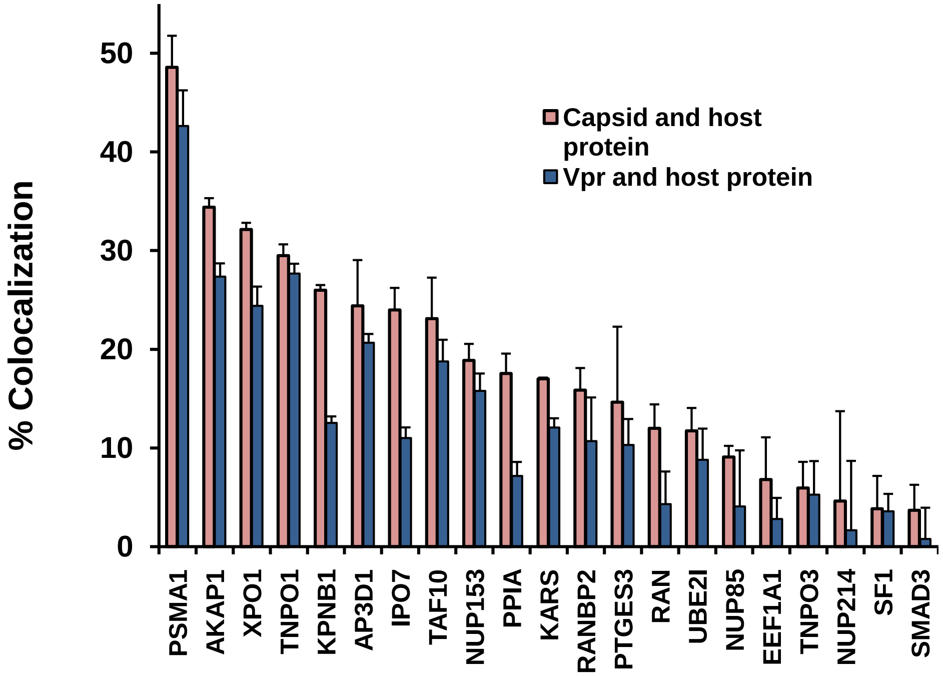
Table 5.10: Enriched GO molecular processes in the unique set of human proteins that involve novel and Group 2 predicted interactions.

GO term ID	GO term name	p-value
GO:0032991	macromolecular complex	2.07e-45
GO:0031974	membrane-enclosed lumen	2.76e-35
GO:0005886	plasma membrane	1.36e-24
GO:0042611	MHC protein complex	1.26e-18
GO:0005829	cytosol	1.30e-17
GO:0043226	organelle	1.19e-13
GO:0044459	plasma membrane part	1.68e-13
GO:0005634	nucleus	2.11e-10
GO:0043234	protein complex	6.48e-10
GO:0005575	cellular-component	1.47e-08

Table 5.11: Enriched GO cellular components in the unique set of human proteins that involve novel and Group 2 predicted interactions.

5.3.5 Colocalization of Selected Host Proteins with Vpr and Capsid

As a follow up experiments, a subset of host proteins that were predicted to interact with an HIV-1 protein according to the model were selected to check whether they colocalize with the viral proteins vpr and capsid. The experiments were conducted through cell live imaging microscopy (carried out by external collaborators Section 5.5). As proximity of two proteins is not enough evidence to conclude two proteins are in contact, these colocalization experiments cannot validate the presence of an interaction. However, they provide additional support for the validity of the interaction and the fact that the experiments were carried out in living cells strength this support. Percent colocalization results of 21 host proteins with vpr and capsid are shown in Figure 5.7. In these experiments, cyclophilin A PPIA or CypA) is used as a positive control as it is known to interact both with vpr [176] and capsid [177]. If one assumes that every pair that has percent colocalization $> 0\%$ is colocalized, then 4 predictions for vpr are confirmed. These are PPIA (CypA), KARS, EEF1A1 and SMAD3 (see Table 5.12). KARS and PPIA(CypA) showed stronger signal than EEF1A1 and SMAD3. Earlier *in vitro* pull down assays, KARS has been reported to interact with capsid [178]. Table 5.12 provides more detailed information on these four proteins. Similarly, among the 21 host proteins, there are 4 proteins that are identified as predicted interaction partners. Those 4 proteins are PTGES3, SMAD3, UBE2I and XPO1. These four proteins showed a colocalization to varying degrees Table 5.13. In the infectivity assays described in section above, PTGES3's role as a host factor is rejected and XPO1 was classified into the indeterminate set. It could be that the colocalization of these two proteins is unspecific. SMAD3, on the other hand, shows a weak colocalization result.



	PSMA1	AKAP1	XPO1	TNPO1	KPNB1	AP3D1	IPO7	TAF10	NUP153	PPIA	KARS	RANBP2	PTGES3	RAN	UBE2I	NUP85	EEF1A1	TNPO3	NUP214	SF1	SMAD3
Predicted to interact with an HIV-1 protein	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓		✓		✓	✓	✓
Predicted to interact with capsid										✓	✓						✓				✓
Predicted to interact with vpr			✓										✓		✓						✓
Reported in NIAID with an HIV-1 protein	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓		✓
Reported in NIAID with capsid							✓			✓	✓										
Reported in NIAID with vpr			✓	✓	✓		✓			✓	✓		✓			✓					
Konig et al. RNA interference screen hit	✓				✓				✓		✓	✓	✓					✓	✓		
Brass et al. RNA interference screen hit																✓		✓			
Zhou et al. RNA interference screen hit			✓																		
Yeung et al. RNA interference screen hit																					

Figure 5.7: Percent colocalization of host proteins with vpr(blue) and capsid(blue). The percent colocalization is calculated by number of particles overlapping divided by the number of viral protein particles. The experiments are averaged over 3 replicates. The first row indicates whether the host factor is predicted to interact with any of the viral proteins; those that are predicted are shown in green and check marked. The next two rows in the lower table list whether the corresponding protein is predicted with capsid or vpr. In the subsequent rows, the information whether the host protein is called as a host factor in the four genome-wide screens is given. Whether in König et al. [74], Brass et al. [72], Zhou et al. [73] or Yeung et al. [79], the white cells indicate the human genes are not called as a host factor by the screen.

Gene	Gene name	%Colocalization with capsid(± std)	Evidences	HIV-1 interactors reported in NIAID (keywords)	Normalized RF score
PPIA	cyclophilin A	17.54(1.66)	detected in virion	nef[binds], vif[binds], vpr[isomerized by], gp120[inhibits, requires], capsid[binds, interacts with, isomerized by, modulates, stabilizes], matrix[binds], pr55[binds, incorporates, modulated by]	0.42
KARS	lysyl-tRNA synthetase	17.01(3.20)	confirmed, König	vpr[inhibits], capsid[interacts with], pr55[incorporates, interacts with], protease[inhibited by]	0.32
EEF1A1	eukaryotic translation elongation factor 1 alpha 1	6.79(0.51)	-	tat[interacts with], matrix[binds, inhibits], integrase[binds]	0.02
SMAD3	SMAD family member 3	3.69(0.92)	-	tat[inhibited by,stimulated by]	0.02

Table 5.12: Colocalization results for genes predicted to interact with capsid.

Gene	Gene name	%Colocalization with vpr(± std)	Evidences	HIV-1 interactors reported in NIAID (keywords)	Normalized RF score
PTGES3	Prostaglandin E synthase 3	10.30(0.95)	rejected, König	vpr[interacts with]	0.34
SMAD3	SMAD family member 3	0.77(1.33)	N/A	tat[inhibited by, stimulated by]	0.33
UBE2I	ubiquitin-conjugating enzyme E2I	8.80(1.95)	N/A	p6[interacts with]	0.22
XPO1	Exportin 1	24.40(3.39)	indeterminate, Zhou	rev[binds, co-localizes with, enhanced by, exported by, inhibited by, interacts with, modulated by, recruits], vpr[exported by], matrix[exported by]	0.01

Table 5.13: Colocalization results for genes predicted to interact with vpr.

5.4 Conclusions

In order to assist the experimental efforts for rapidly inferring the complete HIV-1,human physical interactome a supervised model for predicting HIV-1, human PPI prediction task was presented in this Chapter. This model was to my knowledge the first attempt in the literature that predicted the global set of interactions between HIV-1 and human host cellular proteins [125]. In building this model, relevant biological information is compiled from various data sources and encoded as features to serve as evidence in the prediction task. The biological information included co-occurrence of functional motifs and their interaction domains and protein classes, gene ontology annotations, posttranslational modifications, tissue distributions and gene expression profiles, topological properties of the human protein in the interaction network and the similarity of HIV-1 proteins to human proteins' known binding partners. A Random Forest classifier was built on this extensive dataset. The model's predictions achieved an average Mean Average Precision (MAP) score of 23%. Gene Ontology enrichment analysis in the predicted set of human partners identified immune system related processes as the top ranked molecular processes. In Chapter 5 we provided a list of putative interactions based on a supervised learning model. Pittsburgh Center for HIV-1 Protein Interactions conducted live cell imaging experiments to check 21 host proteins' colocalization with vpr and capsid. These results provided additional evidence in support of some of our predictions.

5.5 External Collaborators

Colocalization experiments were conducted by the laboratory Dr. Simon Watkins of University of Pittsburgh as part of the Pittsburgh Center for HIV-1 Protein Interactions.

Chapter 6

Extended Model with New Feature Set

6.1 Overview

In Chapter 5, a supervised classifier for predicting HIV-1,host physical protein-protein interactions was introduced. This model was trained on 35 features derived from various biological data sources and information specific to HIV-1 viral infection. These features carried indirect information about PPIs such as gene expression of host proteins in the presence of HIV-1 infection, domain and sequence pairs, functional annotations of the proteins and relationship of human proteins with one another. Since then, several new biological information sources pertinent to HIV-1, host relation have become available. This also includes for the first time the direct interaction data. Here, we present an improved model into which we incorporated these additional sources. These datasets include i) four genome-wide RNA inference screens aimed at establishing host proteins required for HIV-1 replication, ii) a high-throughput immunoprecipitation pull-down mass spectrometry assay for detecting viral and host PPIs, iii) a set of human proteins reported to interact with viruses' proteins and iv) knowledge of human proteins detected in budding HIV-1 virions. We utilized our knowledge of the context of the host cellular machinery to incorporate new features into our models based on these datasets. The model trained with the richer feature set outperforms the first model by a 13% relative increase in mean average precision. The set of predicted interactions serves as biological hypotheses to test. In this section, these biological information sources and features derived from them are detailed; subsequently, we report empirical performance evaluation of the model.

6.2 Methods

6.2.1 Problem Setting and Formulation

The same problem formulation as described in Chapter 5 for the earlier model was applied to develop the new model described in this chapter. Predicting physical interactions between HIV-1 and human protein pairs is cast as a binary classification task. That is, each viral-human protein pair belongs to one of two classes: 'interaction' or 'non-interaction'. Every protein is described with a numeric feature vector, where each feature describes properties of the human and viral proteins. Using labeled examples of the two classes and the feature vectors, a function that distinguishes the two classes is learned. Again, a supervised classifier, the Random Forest classifier [145], was used (see Section 5.2.2.2 for details).

The positive and negative interactions employed were the same as in the previous model presented in Chapter 5. The positive interactions are the group 1 subset of HIV-1, host PPIs deposited in the NIAID database [134, 135]. The positive set included 1063 proteins involving 17 HIV-1 proteins and 721 human proteins. The negative dataset was chosen uniformly at random. The ratio of the negative to positive class was assumed to be 100:1, a value chosen based on the average number of such interactions involving HIV-1 proteins. In training, the negative to positive example ratio was treated as a parameter and was optimized through cross-validation.

6.2.2 New Features Included

Feature name	**Number of features**	**Feature type**	**Description**
IP-MS spectrometry	1	HV, HPPI	Encodes the score of the protein pair in the high-throughput IP mass spectrometry pull down assay.
IP-MS spectrometry complex	1	HV, HPPI	Encodes the IP-Mass spectrometry results together with human complex information.
RNAi hits	1	H	Encodes in fraction of genome wide RNAi screens the gene is called as a hit.
RNAi hits complex enrichment	1	H, HPPI	Encodes whether the human protein is part of an RNAi hit enriched complex.
RNAi hits pathway enrichment	1	H, HPPI	Encodes whether the human protein is part of an RNAi hit enriched pathway.
In virion feature	1	H	Encodes whether the human protein is detected in a budding virion.
Other virus interactors	1	H	Encodes if the human proteins is reported to interact with other human infecting viruses.

Table 6.1: Additional features derived for prediction of interactions between HIV-1 and human proteins. The first column lists the name of feature; the second column indicates number of features of this type. The third column, 'feature type', describes whether the feature is specific to the HIV-1 protein pair (HV), specific only to the human protein (H) or specific only to the HIV-1 protein (V). Those features that make use of the human proteins connectivity knowledge are also indicated in this column by 'HPPI'. The last column provides a brief description of how the feature was encoded.

The previous feature set used to train the first model described in Chapter 5 to predict PPIs between HIV-1, human proteins, was derived from various data sources. A total of 35 features were derived. This feature set is summarized in Table 5.4 and details of the data sources and the encodings are provided in Section 5.2.4. This feature set will referred to as *feature set 1*. Feature set 1 was expanded by the addition of new features, which are listed and summarized in Table 6.1. This expanded feature set includes a total of 42 features and will be referred to as the *final feature set*. Below, the new features and their encodings are detailed and the rationale behind them are explained.

6.2.2.1 Features Based on Interaction Data Derived from Immunoprecipitation Assay Coupled with Mass Spectrometry

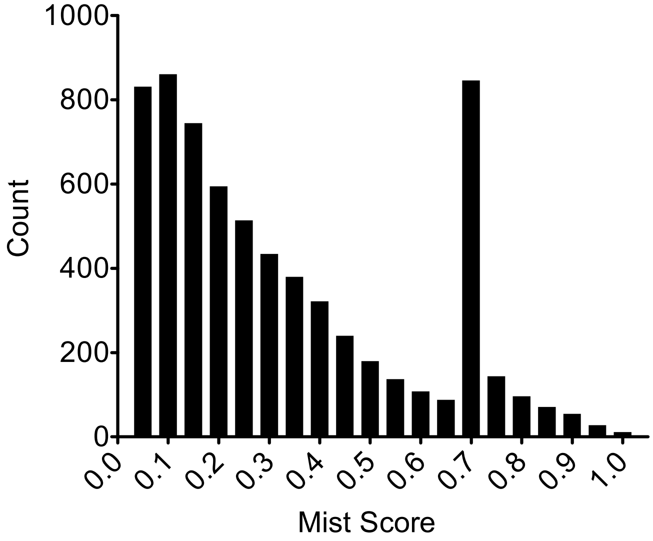


Figure 6.1: The HIV-1, human protein pairs identified in the IP-MS experiment are scored according to specificity, reproducibility and abundance.

Immunoprecipitation coupled with mass spectrometry (IP-MS) identification is a powerful method for detecting PPIs *in vitro*, as discussed in Section 2.2.1.2. Krogan et al. applied this method to identify complexes formed between HIV-1 proteins and human proteins (unpublished). A list of human proteins that were identified in pull-down experiments from HEK293 cells transfected with single HIV-1 proteins was obtained from the Krogan lab. In this study, every HIV-1 protein was analyzed at least four times in independent pull-down experiments; the resulting HIV-1, human protein pairs were scored according to specificity, reproducibility and abundance referred to as MIST score (Krogan et al. unpublished). Figure 6.1 displays the distribution of these scores. The higher the score, the more likely the protein pair represents a true interaction. We encoded two features using this dataset:

Mass spectrometry feature: The first mass spectrometry feature for human protein h and viral protein v ,$f_{\text{mass}}(h,v)$, was encoded based on discretization of the MIST score into four bins over the range [-1,1]. Protein pairs that were not pulled down were scored as -1.

Mass spectrometry complex feature: IP-MS experiments suffer from certain limitations and biases as described in Section 2.2.1.2. The primary problem of these experiments for their use in PPI prediction task is inclusion of indirect interaction partners. To account for such possible experimental bias, a second feature was included, in which these indirect relationships were taken into account. For each of the human proteins pulled down in the mass spectrometry experiment I checked if, other human proteins which participate in a complex with the respective protein. The complex membership information was retrieved from the CORUM database (see below). The proteins pulled down by a given viral protein and the complexed human proteins were scored in the same way as the actually identified human-viral protein pair.

Specifically, let H be the set of human proteins pulled down with virus protein v and let M be the set of human proteins known to be in a complex with at least one human protein in H. The mass spectrometry complex feature (masscomplex) $\forall$ m in M is defined as:

$$f_{\text{masscomplex}}(m, v) = \max_{h \in H} \left(f_{\text{mass}}(h, v) \right) \tag{6.1}$$

The human protein complexes were obtained from CORUM database, downloaded on 08/19/2009 [179]. The dataset contains 1,342 human protein complexes, which were manually annotated from experiments published in scientific articles.

6.2.2.2 RNA Interference Screens Hits

Four genome-wide RNAi screens have been conducted with HIV-1 to identify host factors required for viral infection, as listed in Table 2.5. Although these screens aimed at identifying the complete set of host factors, the resulting gene sets lacked overlap (see Figure 6.3 a). While the individual genes identified by the RNAi screens may not be reproducible between screens, the screens identified related genes. For instance, many of the subunits of the mediator complex were identified by one of the screens but only a few of them were identified by multiple screens (Figure 6.2). To quantify this observation, we calculated the overlap among the screens by taking into account the cellular interactions within the host. Two genes were considered overlapping if they were part of the same complex (see Figure 6.3 or b) or part of the same pathway (see Figure 6.3 c). With this modified criterion, the pairwise overlaps among the screens are much larger (compare Figure 6.3 a with 6.3 b and c). These results indicate that host factors detected in different screens are indirect or direct interactors of one another. When creating RNAi features, these indirect relationships were considered. Below, we describe the three features developed based on the RNAi screens.

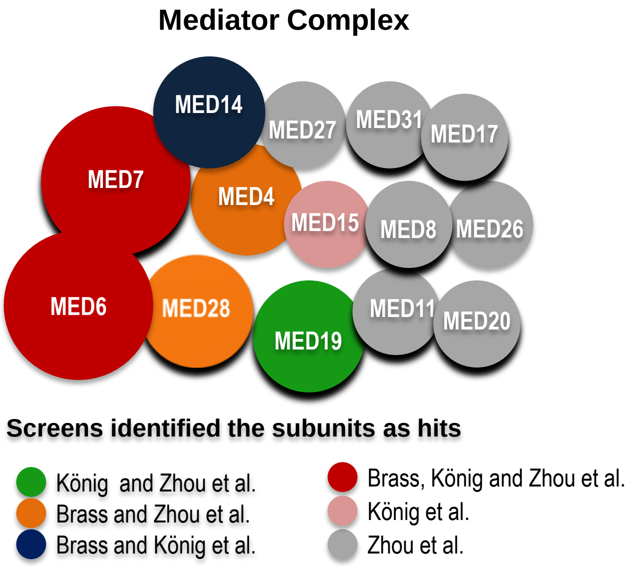
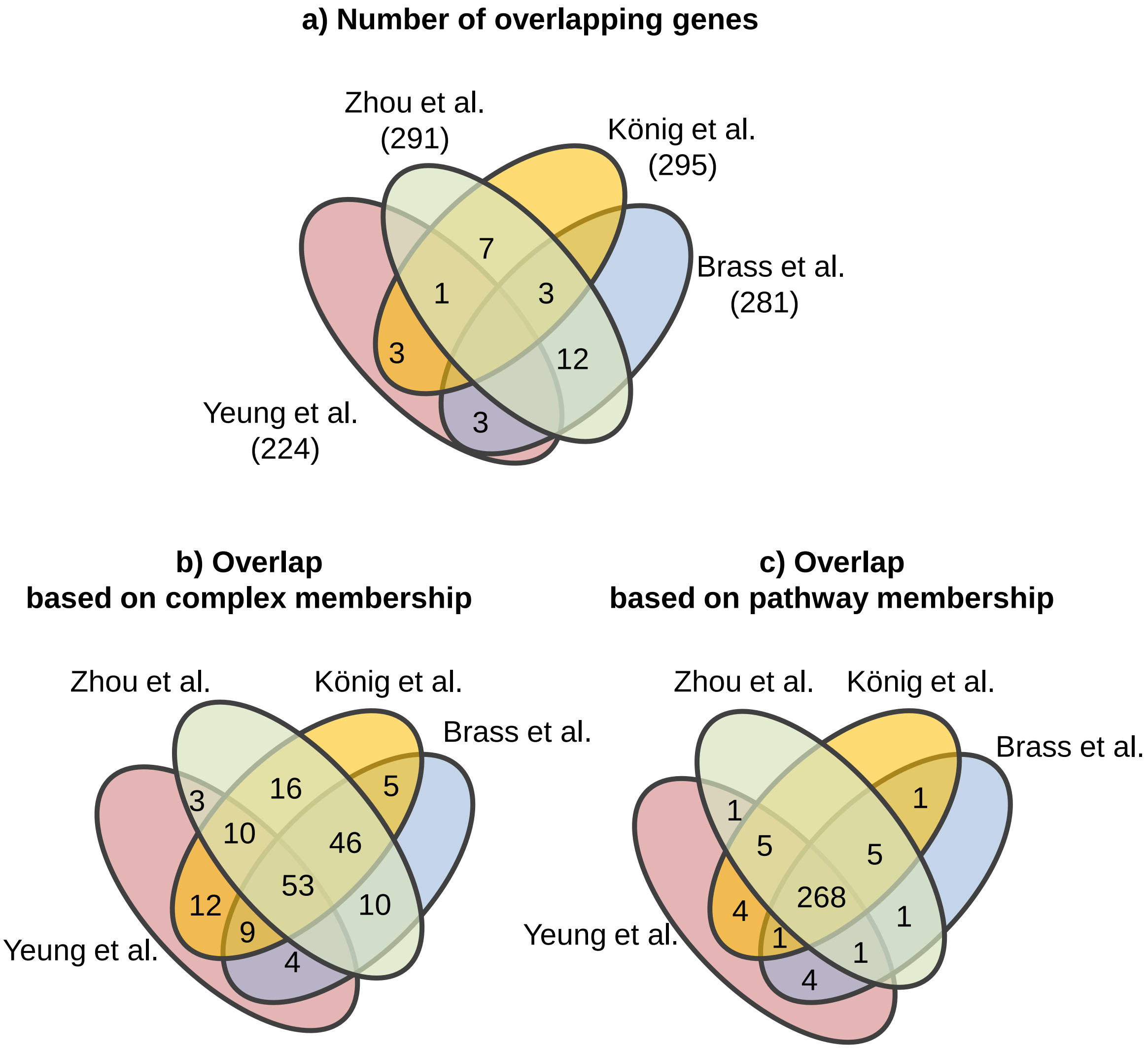


Figure 6.2: Mediator complex subunits identified by one or more genome-wide RNA inferences screens.

RNA interference hits feature: This feature scores the viral-host protein pairs based on how many screens called the human gene as a host factor important for the viral infection.



a) Number of overlapping genes

b) Overlap based on complex membership

c) Overlap based on pathway membership

Figure 6.3: Overlap among the four published genome-wide RNA inference screens based on different overlapping criteria. Overlap is calculated if genes in different screens are a) identical, b) share a complex membership and c) share a pathway membership.

RNA interference complex enrichment features: Because the genes identified in the RNAi screens do not necessarily detect directly interacting host factors and my analysis on the overlap shows that different screens capture related proteins, I incorporated the cellular context. This feature utilizes the human protein complexes that are statistically enriched by the RNAi hit genes. Examples of human protein complexes which include statistically significant numbers of RNAi detected genes are given in Table 6.2. The feature scores host viral protein pairs that belong to one of the enriched complexes as 1 and as -1 otherwise. The human protein complexes were obtained from CORUM database (downloaded on 08/19/2009)[179].

RNA interference pathway enrichment: Similar to the complex enrichment feature, we make use of the pathway knowledge in this feature. Compared to protein complexes, pathways include more distant, indirect relationships. We identified human protein pathways that are statistically enriched with RNAi hits (p-value ≤ 0.01). The human proteins part of these pathways formed a set, P. The feature scores the protein pairs, which are composed of human proteins from this list as 1 and -1 otherwise. Example of enriched pathways are shown in Table 6.3. The human protein pathways were acquired from Pathway Commons, downloaded on 01/26/2010 [180]. Pathways including HIV-1 proteins, or related to HIV-1 infection were excluded from the list of pathways.

Name of the complex	**Number of subunits in the complex**	**Number of genes detected in the RNAi screens and encodes the complex's subunits**	**p-value**
Mediator complex	32	14	8.82e-11
Spliceosome	140	28	1.74e-10
BRCA1-RNA polymerase II complex	26	12	9.30e-10
Nuclear pore complex	28	12	2.67e-09
RNA polymerase II holoenzyme complex	24	11	5.20e-09
PA700-20S-PA28 complex	36	13	7.12e-09
26S proteasome	22	10	2.91e-08
C complex spliceosome	79	17	2.18e-07
CRSP complex	12	7	4.27e-07

Table 6.2: Example of protein complexes enriched in the set of RNA interference detected human genes.

Name of the pathway	**Number of genes in the pathway**	**Number of genes detected in the RNAi screens and encodes the pathway proteins**	**p-value**
Gene expression	360	72	5.36e-25
mRNA capping	276	54	1.31e-18
RNA polymerase II transcription	292	55	3.77e-18
mRNA processing	278	53	8.91e-18
Influenza infection	186	38	4.81e-14
Ubiquitin mediated degradation of phosphorylated Cdc25A	57	17	1.05e-09
mRNA editing: C to U conversion	178	27	1.73e-07

Table 6.3: Example of pathways enriched with RNAi genes.

The statistically enriched complexes and pathways are selected by calculating a p-value of the overlap between the RNAi screens and the pathway/complex genes. For a given pathway or complex, p, we define the set of genes that encode the proteins in the complex or pathway as G_p. The union set of genes called by the four RNAi screens is referred to as G_r. The size of overlap among the two gene sets is k. G_a denotes the set of all possible genes. We use hypergeometric distribution, also known as the one-tailed Fisher's exact test, to calculate a p-value for the size of the overlap:

$$p = \sum_{i=k}^{min(G_p,G_r)} \frac{\binom{G_p}{i}\binom{G_a-G_p}{G_r-i}}{\binom{G_a}{G_r}}$$

6.2.2.3 Human Proteins Detected in HIV-1 Virions

During budding from the cell membrane multiple host proteins, along with genomic RNA and viral proteins, are packaged into the virion. While some of these proteins are merely bystanders and are incorporated into the virion by chance due to their proximity to the virus budding site, others are known to play key roles in the viral life cycle or in pathogenesis [175]. In Chapter 5, we have used this biological information as an external validation of the trained model. Here, we include these data as a feature. The dataset used [175] includes 314 human genes detected experimentally in the virion. The virus-host protein pairs were scored as 1 if the human protein belongs to this dataset and -1 otherwise.

6.2.2.4 Other Virus-Host Interactions

Protein name	Viruses
Histone acetyltransferase p300	BPV, HAdV, HIV, HPV, MPyV, SV
Cell division control protein 2 homologue	EBV, HAdV, HIV, HPV, SV
Serine/threonine-protein phosphatase 2A	HIV, HPV, MPyV, RSV, SV
CREB-binding protein	HAdV, HIV, HPV, MPyV, SV
TATA-box-binding protein	BPV, HAdV, HIV, HPV, SV
Transcription initiation factor IIB	BPV, HHV, HIV, HPV, SV
Transcription initiation factor TFIID	SV, HIV, HAdV, BPV, HPV

Table 6.4: Example of proteins targeted by multiple viruses. The abbreviations for the viruses are listed in Notation section.

Host-virus interactions involving many different viruses including but not limited to HIV-1 are being catalogued in the VirusMint database [181]. Given that all viruses encounter similar barriers when using a human cell as a host, viruses are likely to recruit similar host cell components. It has been shown that different viruses target proteins that participate in the same pathway [182]. Table 6.4 provides some examples on how the same protein has been reported to interact with different viruses. Capitalizing on these observations, we include the information whether the human protein is targeted by a virus other than HIV-1 as a feature.

6.2.3 Performance Evaluation

The performance of the models was evaluated through 3-fold cross validation experiments. Each cross-validation experiment was repeated 10 times; in each experiment, the negative interaction data and the splits are chosen randomly. The same fold splits were used as with the experiments described in Chapter 5 to allow for a fair comparison. We evaluated the quality of the new predictive model using two figures of merits that ignore the success on the true negative (TN) rate: the receiver operating characteristic (ROC) curve and the precision vs. recall curve [170]. The Mean Average Precision (MAP) score was employed to summarize the precision vs. recall curve and the area under the ROC in order (AUC) to summarize the ROC curve as a scalar score that ranges between 0 and 1 [170]. Since the low false positive (FP) region of the ROC curve is of particular interest in the PPI prediction task, the partial AUC scores AUC50, AUC100, AUC200 and AUC300 were determined by measuring the area under the ROC curve until reaching 50, 100, 200 and 300 FP predictions, respectively.

6.2.4 Gene Ontology Enrichment Analysis

The significantly enriched GO terms in the list of human proteins involved in the predicted interactions were identified using Ontologizer 2.0 [171] using the child-term parent intersection method and using Bonferroni correction for multiple hypothesis testing.

6.3 Results and Discussion

	MAP	**AUC**	**AUC50**	**AUC100**	**AUC200**	**AUC300**
Feature set 1	0.2300	0.9150	0.0670	0.1073	0.1682	0.2156
Final feature set	0.2598	0.9270	0.0797	0.1264	0.1921	0.2424

Table 6.5: Performance comparison of the HIV-1 model trained with feature set 1 and the final feature set. Averages of MAP, AUC and partial AUC scores over 10 repeated 3 fold cross-validation experiments are presented. Standard errors of the experiments can be found in the extended table Table 6.6.

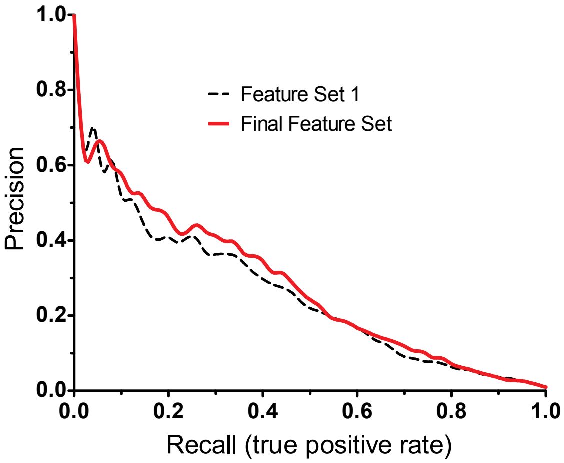


Figure 6.4: Comparison between the two models i) the model trained on feature set 1 decribed in Chapter 5 (red curve) ii) the new model trained on final feature set (black dashed curve).

A Random Forest classifier with this extended feature set (features listed in Table 5.4 and the new features listed in Table 6.1) was trained to contain a total of 42 features. The performance of the model is evaluated through 3 fold cross-validation experiments. Each cross-validation experiment was repeated 10 times; in each experiment the negative interaction data and the data splits were selected randomly. Average precision vs. recall curves of these experiments are shown in Figure 6.4 (red curve). Table 6.5 lists the average MAP, AUC and partial AUC scores of the new model. The new model achieved an average MAP score of 0.26. This is a 13% relative increase on the first model described.

6.3.1 Features' Contribution to the Performance

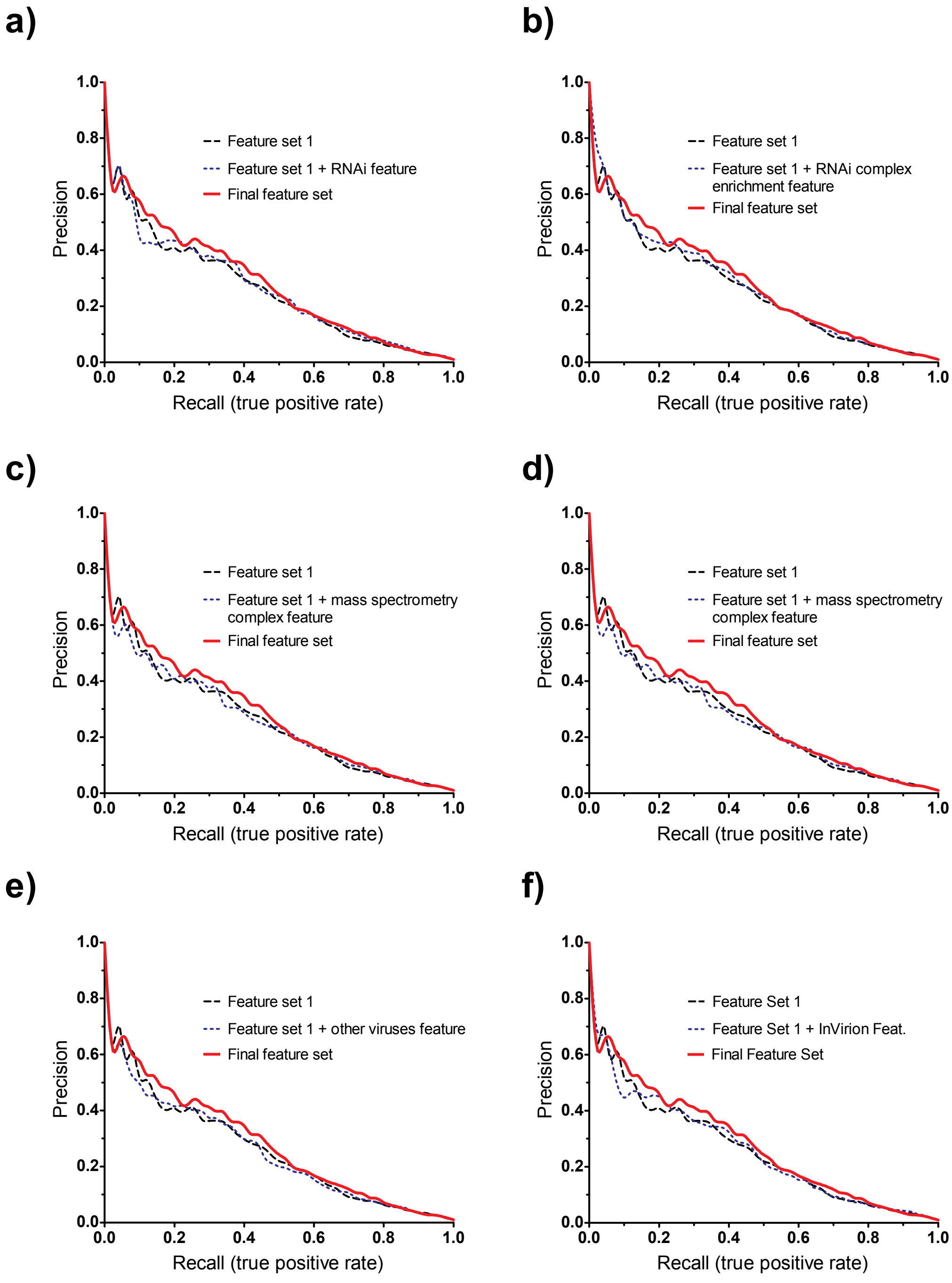
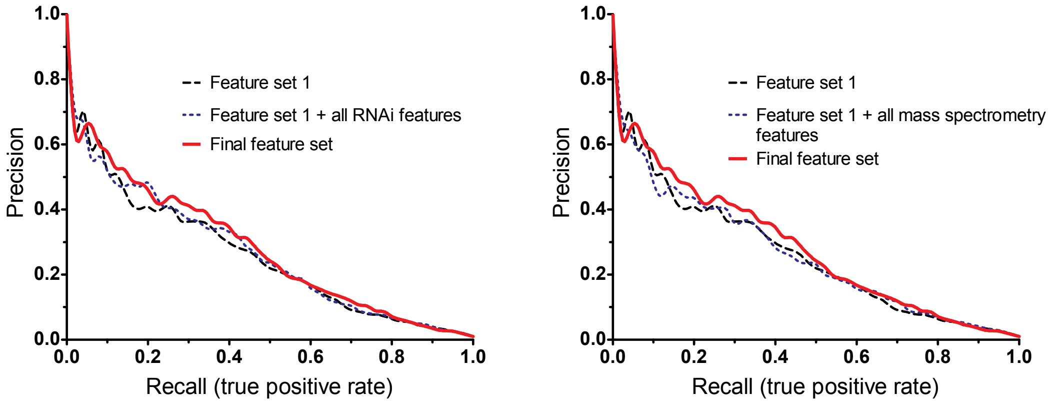


Figure 6.5: Each subfigure shows precision recall curves of three models trained with three different feature sets : i) feature set 1 described in Section 5.2.4 ii) a new feature added to the feature set 1 iii) all new features added to feature set 1.

In order to evaluate the contributions of the individual features to the performance, I trained and tested several models, where new features were added one by one to feature set 1. Figure 6.5 illustrates the precision recall curves of the models trained with feature set 1, adding a new feature to the feature set and the final set where all the new features are added to the feature set 1. The curves indicate that features act differently across different recall ranges. In high recall regions, the mass spectrometry data improve the precision (Figures 6.5 c,d). The RNAi pathway enrichment feature improves the precision in most recall ranges (Figure 6.5 b). The performance comparison based on AUC and MAP scores shows that the in virion features and mass spectrometry features alone did not lead to a significant improvement in MAP and AUC scores (Figures 6.5 e, f). Instead, the RNAi enrichment features lead to the largest increase in performance.



(a) Addition of all RNAi features

(b) Addition of all mass spectrometry features

Figure 6.6: Precision recall curves of three models: i) Model trained with feature set 1 (described in Section 5.2.4) ii) a) all mass spectrometry features b) all RNAi features added to the feature set 1 iii) all new features added to feature set 1(described in Section 6.2.2).

In order to further assess the contribution of RNAi features and mass spectrometry features, the features were grouped and new models were trained. Figure 6.6 demonstrates the performance of these models, pointing to the predictive power of the RNAi features. MAP and AUC scores of these experiments are provided in Table 6.6.

6.3.2 Predicted Interactome

Having evaluated the model, a final model was trained with all available positive interactions. All HIV-1, human pairs were ranked according to their Random Forest score. The score measures the difference between positive and negative votes of the decision trees in the forest. The higher the score of a pair, the higher the likelihood the model predicts that pair to be an interaction. At an Random Forest score cutoff of zero, the model predicts 2803 interactions between 1272 human proteins and 17 HIV-1 viral proteins. The set of predicted interactions depends on the chosen Random Forest score threshold; lowering the threshold will increase the true positive (TP) rate at the expense of a higher false positive (FP) rate. In the following sections, first an experimental result verifying one of the predictions is described. Subsequently, enriched GO terms in the predicted host factors are provided.

6.3.3 An Independent Experimental Validation, SUMO2 Interaction with the Viral Protein Integrase

Strong evidence supporting one of the new predictions has recently been reported independently [183]. Interaction of SUMO2 with HIV-1 protein integrase is in the new prediction list (score 0.15) and has not been reported in the NIAID database. Viral protein integrase takes role in the integration of the viral cDNA into the host genome, and also functions in other steps of replication. SUMO2 is one of the four SUMO proteins that the human genome encodes. Sumoylation is a post-translational modification that consists of the covalent attachment of small ubiquitin-like modifier (SUMO) peptides to a lysine residue. SUMO modification can lead to diverse cellular consequences; it can affect signal transduction, protein stability and localization and transcriptional regulation [184]. It is also well established that several viral proteins are either sumoylated or interfere with the sumoylation pathway [185]. In the case of HIV-1, p6 has been shown to interact with the SUMO1 protein [186], but the interaction with integrase has not been reported.

Zamborlini et al. [183] very recently reported that HIV-1 integrase is sumoylated and that the three lysine residues of integrase sequence are SUMO-acceptor sites. They reported that mutation of SUMO-acceptor residues in integrase led to reduced infectivity and slower replication kinetics. Their experimental results demonstrated that sumoylation-defective integrase mutants showed a significant decrease in integration events compared to HIV-wild type infected cells. These results are also in accordance with the genome-wide siRNA screen of König et al. [74], who also discovered SUMO2 as a required host factor for viral replication. This independent validation shows how several lines of evidences can lead to promising biological hypotheses.

6.3.4 Enriched Functions and Biological Processes

To gain insight into the system properties of the new model, the significantly enriched GO functions, molecular processes and cellular components among the list of predicted human interaction partners were calculated. Terms related to protein binding, transcription regulation activity and nucleic acid binding were highly enriched. Also, proteasome complex terms were highly enriched. The complete list of significantly enriched terms for the three GO ontologies can be found in Tables 6.7, 6.8 and 6.9, respectively. The enriched functions and biological process for this model includes similiar terms as the previous model, such as transcription factor binding but also include GO terms that are different such as muscle contraction or DNA clamp loader activity.

GO term ID	GO term name	p-value
GO:0032395	MHC class II receptor activity	0.00e+00
GO:0003689	DNA clamp loader activity	0.00e+00
GO:0033170	protein-DNA loading ATPase activity	0.00e+00
GO:0046965	retinoid X receptor binding	0.00e+00
GO:0004972	N-methyl-D-aspartate selective glutamate receptor activity	0.00e+00
GO:0005515	protein binding	1.01e-179
GO:0030528	transcription regulator activity	2.56e-61
GO:0003677	DNA binding	5.42e-59
GO:0008134	transcription factor binding	5.77e-59
GO:0003712	transcription cofactor activity	9.16e-55
GO:0000988	protein binding transcription factor activity	1.90e-54

Table 6.7: Examples of enriched GO molecular function terms in the set of human proteins that participate in the predicted HIV-1, host PPI network.

GO term ID	GO term name	p-value
GO:0006936	muscle contraction	8.98e-03
GO:0007229	integrin-mediated signaling pathway	8.94e-03
GO:0043506	regulation of JUN kinase activity	8.94e-03
GO:0071479	cellular response to ionizing radiation	8.77e-03
GO:0045736	negative regulation of cyclin-dependent protein kinase activity	8.77e-03
GO:0060444	branching involved in mammary gland duct morphogenesis	8.77e-03
GO:0007276	gamete generation	8.72e-03
GO:0045137	development of primary sexual characteristics	8.37e-03
GO:0046627	negative regulation of insulin receptor signaling pathway	8.16e-03
GO:0008584	male gonad development	8.12e-03
GO:0031668	cellular response to extracellular stimulus	7.76e-03
GO:0007267	cell-cell signaling	7.68e-03
GO:0046883	regulation of hormone secretion	7.57e-03
GO:0071445	cellular response to protein stimulus	7.37e-03
GO:0006874	cellular calcium ion homeostasis	7.11e-03

Table 6.8: Examples of enriched GO biological process terms in the set of human proteins that participate in the predicted HIV-1, host PPI network.

GO term ID	GO term name	p-value
GO:0005838	proteasome regulatory particle	0.00e+00
GO:0022624	proteasome accessory complex	0.00e+00
GO:0005663	DNA replication factor C complex	0.00e+00
GO:0032991	macromolecular complex	4.21e-174
GO:0005654	nucleoplasm	1.17e-152
GO:0031981	nuclear lumen	1.77e-141
GO:0044428	nuclear part	2.26e-136
GO:0043233	organelle lumen	1.19e-133
GO:0031974	membrane-enclosed lumen	6.63e-133
GO:0005634	nucleus	6.58e-119

Table 6.9: Examples of enriched GO cellular component terms in the set of human proteins that participate in the predicted HIV-1, host PPI network.

6.3.5 Comparison of the First Model and the Old Model

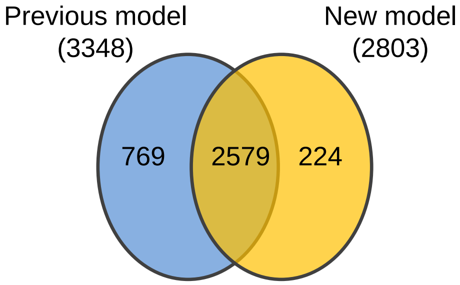


Figure 6.7: Comparison between the two lists of predictions at Random Forest cutoff 0 i) the previous model trained on feature set 1 (decribed in Chapter 5 and ii) the new model trained on final feature set.

A comparison of the two lists of predictions by the previous model trained on feature set 1 and the new model trained on the final feature set is provided in Figure 6.7. Although the overlap between the two predicted sets is quite large (2579), the second model predicts fewer interactions (2803) than the previous model (3803). The simplest interpreation of this finding is that the new information included in the model leads classification of many pairs as negative interactions.

6.4 Conclusions

In this chapter, I described an extended set of features leading to an improved version of the HIV-1, human PPI prediction model described in Chapter 5. The new model method was also based on a supervised learning framework. New biological datasets were incorporated in the model. These datasets included four RNAi screens and interaction data from a high-throughput IP-MS experiment, a list of human proteins detected in budding HIV-1 virions and human proteins targeted by other viruses. Evaluation of features indicated that RNAi features were the most predictive ones, while other features contributed to the precision across different recall ranges. One of our predictions between integrase and human protein SUMO2 has been independently validated experimentally and reported in a recent literature [183].

	MAP	**AUC**	**AUC50**	**AUC100**	**AUC200**	**AUC300**
Feature set 1	0.2300(0.0039)	0.9150(0.0022)	0.0670(0.0025)	0.1073(0.0030)	0.1682(0.0036)	0.2156(0.0042)
Feat. set1 + RNAi feature	0.2318(0.0039)	0.9201(0.0016)	0.0676(0.0023)	0.1086(0.0028)	0.1716(0.0034)	0.2183(0.0041)
Feat. set1 + RNAi complex feature	0.2417(0.0039)	0.9208(0.0013)	0.0721(0.0025)	0.1149(0.0032)	0.1792(0.0037)	0.2279(0.0042)
Feat. set1 + RNAi pathway feature	0.2469(0.0038)	0.9206(0.0016)	0.0755(0.0022)	0.1194(0.0028)	0.1838(0.0038)	0.2322(0.0043)
Feat. set1 + all RNAi feature	0.2419(0.0042)	0.9209(0.0015)	0.0719(0.0024)	0.1155(0.0031)	0.1787(0.0039)	0.2270(0.0044)
Feat. set1 + mass spec. feature	0.2337(0.0041)	0.9227(0.0016)	0.0686(0.0025)	0.1090(0.0030)	0.1722(0.0039)	0.2196(0.0046)
Feat. set1 + all mass spec feature	0.2337(0.0041)	0.9218(0.0014)	0.0684(0.0024)	0.1094(0.0031)	0.1720(0.0040)	0.2194(0.0045)
Feat. set1 + other virus interactions feature	0.2319(0.0039)	0.9197(0.0014)	0.0680(0.0025)	0.1097(0.0031)	0.1716(0.0038)	0.2194(0.0043)
Feat. set1 + in virion feature	0.2368(0.0041)	0.9231(0.0013)	0.0677(0.0023)	0.1099(0.0030)	0.1733(0.0038)	0.2207(0.0045)
Feat. set1 + all mass spec. + all RNAi features	0.2551(0.0043)	0.9231(0.0018)	0.0784(0.0026)	0.1249(0.0032)	0.1909(0.0041)	0.2396(0.0043)
Feat. set1 + all mass spec. + all RNAi + other viruses + in virion features (Feat. set final)	0.2598(0.0044)	0.9270(0.0015)	0.0797(0.0025)	0.1264(0.0031)	0.1921(0.0041)	0.2424(0.0046)

Table 6.6: Performances of HIV-1, human Random Forest model using different feature sets. Averages over 3-fold cross-validation experiments repeated 10 times are listed; the standard errors are provided in parentheses. Abbreviations used in the table are as follows: Feat. set 1: feature set 1, the feature set used in model described in Chapter 5, also mass spec.: mass spectrometry.

Chapter 7

Refining Literature Curated Protein-Protein Interactions with Expert Opinions

7.1 Overview

In Chapter 5 and Chapter 6, supervised models for predicting direct HIV-1, human PPIs were presented. These models made use of a subset of HIV-1, human protein interactions deposited in the NIAID database [134, 135]. This database was curated from published scientific articles and reports a mixture of functional associations and physical interactions. Since the goal is to predict direct physical PPIs, I need to distinguish them from functional and indirect associations. I have selected this set based on the keywords describing of the interactions in the database as explained in Chapter 5. In this chapter, efforts to create a higher-quality data set of direct physical interactions of HIV-1 and human PPIs is described and its value is demonstrated by training and testing a new model. Specifically, I i) collected opinions of HIV-1 experts about the interactions reported in the literature ii) formulated a probabilistic framework to assign reliability scores to interactions based on these expert opinions and iii) trained and tested a model using the data labeled by experts. The method presented here is not limited to curated data on HIV-

1 protein interactions, but is applicable in general to other bodies of literature-curated data where it is possible to collect expert opinions. Therefore, this section will present the method for the general case and then show the specific results for the HIV-1 dataset in particular.

7.2 Reliability of Literature Curated HIV-1, Human Protein-Protein Interactions

PPI databases represent tremendous efforts that have been spent on extracting and organizing interactions reported in small-scale experiments [121, 134, 135, 187–191]. Interactions curated from the literature are usually assumed to be of the highest quality available. However, in recent years several concerns have been raised about the validity of this widely accepted assumption [136–142]. The common curation protocol for literature curated PPI databases has been to capture and deposit all interactions that have passed peer review, without further curator judgment [138]. However, not all the published interactions are equally well-supported by experimental evidence. Some interactions have been validated by multiple groups and techniques, and the resulting reliability depends on the proteins studied; other interactions have not been validated in this way. Mackay et al. [136, 137] argued that many reports of PPIs are founded on 'insufficient data' generated by limited strategies. Additionally, Cusick et al. recently reported that many PPIs are supported by single publications [142].

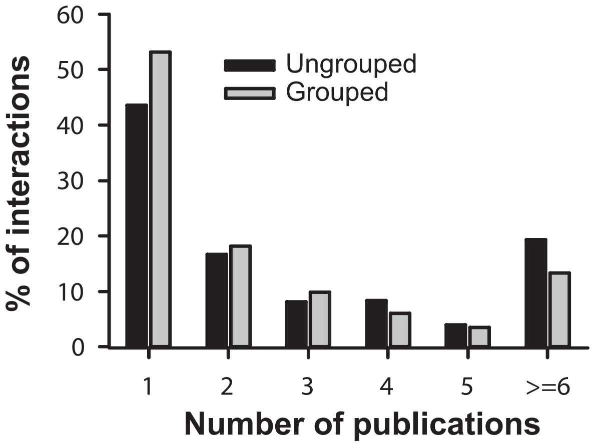


Figure 7.1: Distribution of the number of publications supporting each HIV-1, human protein interaction. The graph depicts two analyses, where publications are counted individually (black) or publications are grouped together if they share a common author (gray). In both cases, the majority of interactions are supported by a single publication. The data of literature curated interactions and the publications reporting them are downloaded from the NIAID HIV-1,human protein interaction database [134, 135].

When I analyzed the number of supporting publications for each of the 2598 HIV-1, human protein interactions in the NIAID database, I found that 44% of all the pairs in the database were reported only in a single publication (see Figure 7.1). When the publications that share at least one common author are grouped together, the statistics become even more striking; the proportion of interactions supported by a group is as high as 53% (see Figure 7.1). The number of interactions that are supported by more than five publications constituted only 19 and 13% of all interactions when publications are ungrouped and grouped, respectively. The lack of follow-up studies-especially by other labs than the one who found the first evidence for an interaction-hints at the possibility that for many of these interactions, there may not be enough experimental evidence to support their inclusion in a gold standard dataset.

Literature curated databases typically present the experimental details of each study in the form of reference to the original publication(s). In theory, this allows the user to review the information and use her own judgment. However, in many systems biology analyses, thousands of interactions are used at a time, which makes it impractical for the investigator to review each piece of experimental evidence at the time of the study. In practice, investigators either assume all small-scale experiments are of equally high quality [192] or disregard some portion of the interactions based on subjective criteria. A more principled approach would be to assign reliability scores for PPIs that reflect the confidence in each interaction. Some databases already implement such reliability scores, e.g. the Molecular Interaction Database (MINT) [138, 193]. MINT's scoring function combines information such as the scale of the experiment, the type of the experiment, the number of publications supporting it and the presence or absence of ortholog interactions. Note, however, that the score is a heuristic and includes several arbitrary parameters [138, 193].

Assessing the data quality of PPIs from small-scale experiments requires a complex judgment of the methods and results of each specific study. Some experimental techniques are more conclusive for identifying functional relations, while others are more robust for direct interactions. Furthermore, the techniques do not work uniformly well across all proteins. In addition to the variability in the powers and limitations of each technique, the conditions under which a study is conducted are important: *in vitro* or *in vivo* environment, the strains used, the mutations introduced, or if there are labels introduced, the attachment sites of the labels. All such parameters should be taken into account when interpreting the results. Such a complex judgment may be provided best by domain experts. Therefore, in order to arrive at reliability scores for the curated HIV-1, human PPIs, I organized a community contribution effort to collect expert opinions. I developed a probabilistic approach to estimate expert labeling accuracies and the reliability scores given to the expert opinions in the absence of benchmark datasets.

7.3 Approach

To obtain HIV-1, human gold standard PPI data, experts were presented with the accumulated published evidence and asked to annotate interacting pairs with labels based on whether they consider the interaction to be supported by enough evidence to conclude that the pair represents a direct physical PPI. Making use of community contribution in solving difficult tasks is a fairly new concept; termed as 'human computation', and it has had several successful applications [194, 195]. The motivation of community contributions is to harness human intelligence for a task that is challenging for computers but for which humans are more capable of and using many human judgments collectively.

Given the task of reviewing PPIs, different experts might have different opinions, especially when there is not enough evidence accumulated in the literature to give a perfectly conclusive answer. Additionally, disagreements among experts might arise because of their biases, expertise and/or stringency levels; e.g., some experts are more difficult to convince with partial evidence or results of certain experimental techniques. For these reasons, expert opinions will be noisy and subjective. By asking several experts about the same interaction pair, the confidence in the interactions can be better assessed. Although having as many expert opinions as possible is beneficial, it was not always possible to obtain several expert opinions on a particular protein pair due to time or expertise constraints. Thus, there was variance in the number of expert opinions for a pair.

Taken these considerations into account, the computational problem becomes the following: given noisy opinions and with possibly inconsistent numbers of judgments for each, how to accurately decide which of the expert-annotated pairs are more likely to have 'direct physical interactions' and the degree of uncertainty of those conclusions. Measuring the uncertainty of a label given multiple expert labels was addressed previously by [196] in an active learning setting. They sought to determine for which data points acquiring labels from other experts would be beneficial; to achieve this, they estimated the uncertainty of the labels given multiple existing noisy labels. Their results showed that repeated labeling can provide additional improvements to predictions, especially when labeling quality is low. However, in their model they assumed that the experts had identical labeling quality, which is unrealistic in many real word problems like is the case for PPIs. In this work, this simplifying assumption is relaxed and do not require experts to have the same labeling quality. Moreover, experts are considered to have different labeling qualities when labeling different class label types, such as 'interaction' class or 'non-interaction' class. In the absence of a ground truth for the labels, given multiple noisy expert opinions the expert labeling accuracies are estimated in each class and these values are used to estimate the most probable label and its associated uncertainty.

In the next section, the problem is formalized and the proposed method to solve this computational problem is described. The effectiveness of the method is demonstrated through synthetic data experiments. Finally, the method was applied to expert opinions on HIV-1, human PPI interactions, which I collected from HIV-1 experts.

7.4 Methods

7.4.1 Problem Formulation

In this section, I describe the problem setting. Assume there are n literature-derived PPIs, on which at least one expert has provided an opinion, and there is no benchmark dataset where true labels are known. Each interaction maps to one true label, y_i, which is unknown and can take one of the two possible label types: $Z = \{z_1, z_2\}$. In this problem, z_1 is the label for 'direct physical interaction' and z_2 is the label for 'not a direct physical interaction'. The set of notations used in this chapter is provided in Table 7.1.

n	the number of interactions about which at least one expert gave an opinion
m	number of experts
Z	set of possible labels
$z_1 \in Z$	label for direct physical interaction
$z_2 \in Z$	label for not a direct physical interaction
$\mathbf{y}$	vector of hidden true labels;where y_i is the true label for interaction i
$\hat{\mathbf{y}}$	matrix of expert opinions; where $\hat{y}_{i,j}$ is the expert j's label for interaction i
$\hat{\mathbf{y}}^*$	vector of estimated labels; $\hat{y}_i^*$ is the most probable label of interaction i
`	expert labeling accuracies, where $\theta_{z,j}$ is the expert j's labeling accuracy for label type z
ˆ	estimated expert labeling accuracies, where $\hat{\theta}_{z,j}$ is the expert j's estimated labeling accuracy for label type z
A_i	set of experts that labeled interaction i
$\gamma_{z_i \times m}$	matrix of probability of labeling an interaction of class type of $z_i \in Z$, where $\gamma_{z_i,j}$ is probability of expert j providing a label for examples in class z_i

Table 7.1: Additional notation used in this chapter.

In the formulation, different expert labeling accuracy across different label types were allowed. This setting is quite realistic as experts may have different error rates in their annotations of 'direct interaction' and 'not direct interaction' classes. There are no benchmark labels and the objective is to find the most probable labels for the interactions given expert labels, $\hat{\mathbf{y}}$:

$$\hat{y}_i^* = \arg\max_{z \in Z} \mathbf{P}\left(y_i = z \mid \hat{\mathbf{y}}_\mathbf{i}, \theta\right) \tag{7.1}$$

The uncertainty of the label type of an interaction, i, is defined as:

$$\hat{u}_i(\hat{y}_i) = 1 - \mathbf{P}\left(y_i = \hat{y}_i^* \mid \hat{\mathbf{y}}_\mathbf{i}, \theta\right) \tag{7.2}$$

Since the set of expert labeling accuracies, θ, is unknown *a priori*, first they were estimated. In the following section the estimation of θ is detailed.

7.4.2 Estimating Experts' Labeling Qualities

The expert labeling qualities were estimated through maximum likelihood estimation (MLE). MLE of expert labeling qualities, $\hat{\theta}^{\text{mle}}$, is the one that maximizes the likelihood of the observed expert opinions:

$$\hat{\theta}^{\text{mle}} = \arg\max_{\theta} \mathcal{L}(\hat{\mathbf{y}} \mid \theta) \tag{7.3}$$

Below, I present how to estimate $\hat{\theta}^{mle}$ for the case in which every interaction receives opinions from every expert; that is, $A_i = 1, \ldots, m$. In the following section, this assumption is relaxed and I handle the cases where it does not necessarily hold.

Case 1: Every expert provide labels for every example (global annotation case)

The log-likelihood of the observed expert opinions can be written as follows:

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{i=1}^{n} \log \mathbf{P}\left(\hat{\mathbf{y}}_\mathbf{i} \mid \theta\right) \\
&= \sum_{i=1}^{n} \log \sum_{z=1}^{2} \mathbf{P}\left(\hat{\mathbf{y}}_\mathbf{i} \mid y_i = z, \theta\right) \mathbf{P}\left(y_i = z\right)
\end{aligned}$$

In the last line, I marginalized over the hidden true label, y_i. It is assumed that decisions by the experts are conditionally independent given the true label. Under this assumption, the log-likelihood can be rewritten as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log \sum_{z=1}^{2} \left(\prod_{j=1}^{m} \mathbf{P}\left(\hat{y}_{i,j} \,|\, y_i = z, \theta\right) \mathbf{P}\left(y_i = z\right) \right) \tag{7.4}$$

In Eq. 7.4, $\mathbf{P}\left(\hat{y}_{i,j} \,|\, y_i = z, \theta\right)$ is the probability of observing expert label $y_{i,j}$ for interaction i, given the true label of that interaction is $y_i = z$:

$$\mathbf{P}\left(\hat{y}_{i,j} | y_i = z, \theta_{z,j}\right) = \theta_{z,j}^{h(\hat{y}_{i,j}=z)} (1 - \theta_{z,j})^{1-h(\hat{y}_{i,j}=z)} \tag{7.5}$$

where h is the indicator function. In Eq. 7.4, $\mathbf{P}\left(y_i = z\right)$ is the prior probability of an interaction belonging to class z; it is assumed that this prior probability is the same for all $i = 1...n$. For this prior probability, an estimate of the class distribution derived from majority vote labels was used. In order to obtain $\hat{\theta}^{mle}$, first Eq. 7.5 is inserted into Eq. 7.4 and next, the expectation-maximization (EM) algorithm [197, 198] is applied to maximize a lower bound of this incomplete data likelihood:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^{n} \log \sum_{z=1}^{2} \mathbf{P}\left(\hat{\mathbf{y}}_i, y_i = z \,|\, \theta\right) \\ &\geq \sum_{i=1}^{n} \sum_{z=1}^{2} \log \mathbf{P}\left(\hat{\mathbf{y}}_\mathbf{i}, y_i = z \mid \theta\right) \\ &= \sum_{i=1}^{n} \sum_{z=1}^{2} g_i(z) \log \frac{\mathbf{P}\left(\hat{\mathbf{y}}_\mathbf{i}, y_i = z \,|\, \theta\right)}{g_i(z)} \end{aligned} \tag{7.6}$$

The first line follows from Jensen's inequality. It is iteratively maximized with respect to the probability distribution $g(z)$ and θ in the expectation and maximization steps, respectively. The derived update equations for step $t + 1$ are as follows:

E-step:

$$\begin{aligned} g_i^{(t+1)}(z') &= \mathbf{P}\left(y_i = z' \mid \hat{\mathbf{y}}_{\mathbf{i}}, \theta^{(t)}\right) \\ &= \frac{\prod_{j=1}^{m} \mathbf{P}\left(\hat{y}_{i,j} \mid y_i = z', \theta^{(t)}\right) \mathbf{P}\left(y_i = z'\right)}{\sum_{z=1}^{2} \mathbf{P}\left(y_i = z\right) \prod_{j=1}^{m} \mathbf{P}\left(\hat{y}_{i,j} \mid y_i = z, \theta^{(t)}\right)} \end{aligned} \tag{7.7}$$

M-step:

$$\theta_{z',j'}^{(t+1)} = \frac{\sum_{i=1}^{n} g_i^{(t)}(z') h(\hat{y}_{i,j'} = z')}{\sum_{i=1}^{n} g_i^{(t)}(z')} \tag{7.8}$$

To obtain $\hat{\theta}^{\mathrm{mle}}$ the procedure is repeated until convergence.

Case 2: Experts only label a subset of examples (subset annotation case)

All experts might not be available or able to annotate every instance due to time and cost limitations, or their expertise might cover only a subset of the examples. Thus, the assumption that every expert labels every instance may not hold. In this section, I provide the solution when this assumption is relaxed. Let the set of labelers of interaction i be a subset, $A_i \subset \{1, \ldots, m\}$. It was required that each interaction has received at least one opinion from at least one expert. In this case, the likelihood function Eq. 7.4 and the EM update equations (Eq. Eq. 7.7 and Eq. 7.8) are modified as follows:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log \sum_{z=1}^{2} \prod_{\substack{j=1 \\ j: j \in A_i}}^{m} \mathbf{P}\left(\hat{y}_{i,j} \mid y_i = z, \theta_{z,j}\right) P(y_i = z) \tag{7.9}$$

Similarly the update equations, Eq. 7.7and Eq. 7.8 are modified:

E-step:

$$g_i^{(t+1)}(z') = \frac{\mathbf{P}\left(y_i = z'\right) \prod\limits_{\substack{j=1 \\ j:j\in A_i}}^{m} \mathbf{P}\left(\hat{y}_{i,j} \mid y_i = z', \theta^{(t)}\right)}{\sum\limits_{z=1}^{2} \mathbf{P}\left(y_i = z\right) \prod\limits_{\substack{j=1 \\ j:j\in A_i}}^{m} \mathbf{P}\left(\hat{y}_{i,j} \mid y_i = z, \theta^{(t)}\right)} \tag{7.10}$$

M-step:

$$\theta_{z',j'}^{(t+1)} = \frac{\sum\limits_{\substack{i=1 \\ i:j'\in A_i}}^{n} g_i^{(t)}(z') h(\hat{y}_{i,j'} = z')}{\sum\limits_{\substack{i=1 \\ i:j'\in A_i}}^{n} g_i^{(t)}(z')} \tag{7.11}$$

Once the expert labeling accuracies are obtained as per the above procedure, they can be plugged into equations 7.1 and 7.2 to find the most probable label and the certainty of that label.

7.4.3 Synthetic Data Experiments

As there were no real data with opinions and expert labels, synthetic data experiments (see Algorithm 2) were carried out to test the effectiveness of our method as follows.

7.4.3.1 Experimental Set Up

Given the prior distribution of classes, true labels were first generated randomly. Meanwhile, each expert's true labeling quality on each class, $\theta_{j,z}$, was assigned uniformly at random by drawing from the interval [0.5-1.0]. The rationale is that experts likely give better-than-random answers. Next, to create an expert's opinion on an instance, the true labels were taken and converted to incorrect label type randomly at the expert's error rate (1-$\theta_{j,z}$).

Algorithm 2 Set up of the synthetic data experiment to estimate expert labeling accuracies

```
**Input:** number of interactions $n$, class priors $\mathbf{P}(y = z)$, for all $z \in \{z_1, z_2\}$, number of runs $N$, $m$ number of experts, probability of labeling an interaction of class type by each expert $\gamma_{z \times m}$
**Output:** maximum likelihood estimate of expert labeling accuracies, $\hat{\theta}^{\text{mle}}_{2 \times z}$, estimated labels $\hat{y}_i$ and their uncertainties $\hat{u}_i$ $\forall i \in \{i = 1, \ldots, n\}$.
**Synthetic experiment:**
Initialize: $r = 0$ and $\forall i$ $A_i = \{\}$
**while** $r \leq N$ **do**
  1. Assign randomly true labels, $\mathcal{Y} = \{y_1, \ldots, y_n\}$ based on $\mathbf{P}(y = z)$
  2. Assign $\theta$ randomly uniform from $[0.5, 1]$
  **for all** data points $i = 1, \ldots, n$ **do**
    **for all** experts $j = 1, \ldots, m$ **do**
      2. True class label, $z_1 = y_i$ and the opposite class label, $z_2 =$ not $y_i$
      3. Pick $\alpha_1$ uniformly at random from [0,1]
      **if** $\gamma_{z_1,j} \leq \alpha_1$ **then**
        Generate expert opinion:
        Pick $\alpha_2$ uniformly at random from [0,1]
        **if** $\alpha_2 \leq \theta_{z_1,j}$ **then**
          Expert $j$ agrees with the true label $\hat{y}_{i,j} = z_1$
        **else**
          Expert $j$ disagrees with the true label $\hat{y}_{i,j} = z_2$
        **end if**
      **end if**
    **end for**
    $A_i = A_i \cup \{j\}$
  **end for**
  $r = r + 1$
**end while**
```

Two scenarios were considered: i) in the *global annotation* scenario, every expert labels every example and ii) in the *subset annotation* scenario, each instance receives labels from a subset of labelers. To achieve this, a probability of labeling an instance to each expert, $\gamma_{j,z}$, was assigned. Each $\gamma_{j,z}$ was drawn uniformly at random from the interval [0,1]. Based on the assigned probability of labeling, the expert gives a set of opinions for a subset of instances; $\gamma_{j,z} = 1$ indicates the expert j labels all of the instances for class z. The flow of the synthetic experiment is given in Algorithm 2. This step is summarized in Algorithm 2.

7.4.3.2 Baseline Estimators

The most probable label estimation was compared to the four following estimators; each labels the interaction as a 'direct interaction' if:

1. the majority of the experts label them as 'direct interaction' (**Majority voting**).
2. there is at least one expert that thinks it is a 'direct interaction' (**Single voting**)
3. there is at least two experts voting for 'direct interaction' (**Double voting**)
4. all the experts agree on the 'direct interaction' label (**All voting**)

7.4.3.3 Evaluation

The synthetic experiments were repeated $N = 300$ times; the mean and standard error of MSE were calculated over these N random configurations. To measure how accurately the maximum likelihood estimator can recover the true expert labeling accuracies and uncertainties, the average mean squared error (AMSE) was calculated:

$$AMSE(\hat{\theta}^{\text{mle}}) = \frac{1}{n}\frac{1}{2m}\sum_{z=1}^{2}\sum_{j=1}^{m}(\hat{\theta}^{\text{mle}}_{z,j} - \theta_{z,j})^2 \tag{7.12}$$

$$AMSE(\hat{u}) = \frac{1}{n}\left(\hat{u}_i(\hat{y}_i, \hat{\theta}^{\text{mle}}) - u_i(\hat{y}_i, \theta)\right)^2 \tag{7.13}$$

In order to assess whether the accuracy of the final label is correctly assigned, we precision and recall rates were reported. The precision is the fraction of the true direct interactions that are identified by the method as 'direct interaction'. On the other hand, recall is the fraction of the correctly identified 'direct interaction' pairs among all the pairs that are direct interactions:

$$\text{precision} = \frac{\sum_{i=1}^{n} h(\hat{y}_i = y_i = 1)}{\sum_{i=1}^{n} h(\hat{y}_i = 1)}, \text{recall} = \frac{\sum_{i=1}^{n} h(\hat{y}_i = y_i = 1)}{\sum_{i=1}^{n} h(y_i = 1)} \tag{7.14}$$

7.4.4 Annotating HIV-1 Human Interaction Labels

I contacted 16 HIV-1 domain experts and requested their opinions about a subset of the interactions deposited in the NIAID database. One of the experts was a PhD student working with HIV-1 experimentally; all others were professors at different universities who have worked several years on one or more HIV-1 proteins. The experts were asked to annotate only the interactions of the HIV-1 proteins on which they consider themselves experts. For each HIV-1 protein, an Excel file was prepared, in which interactions with the HIV-1 protein were listed. The file included the interaction partners of the HIV-1 protein, the keywords retrieved from the NIAID database, and hyperlinks to the original publications so the experts could check the articles if necessary. For HIV-1 proteins, where the number of interactions are ($\leq$ 50), I sent all the interactions reported in the NIAID database regardless of the keyword indicating the association. In cases where >50 interaction partners were listed for HIV-1 proteins, experts were sent only the subset of interactions described with the keywords 'interacts with' and 'binds' in the NIAID database. This was to avoid overwhelming the expert with a long list and to increase the chance of receiving a response. In some cases, experts did not label the files due to time constraints; instead, they provided us with a set of interaction partners they thought were real and direct interactions. Experts were asked to include labels as an additional column in this Excel file: 0 if they thought the reported interaction was *not* a direct interaction, label 1 if they thought it was direct interaction and label 2 if they thought it was indirect or if they were unsure. Upon retrieval of the annotations, I realized experts used different distinctions for indirect interaction and not direct interaction. In order to mitigate this effect, I post-processed the labels to reduce them to two classes: expert-labeled as 'direct interaction' or 'not'.

Of the 2498 interactions, I was able to elicit 765 opinions across 384 interactions. 16 HIV-1 experts contributed labels for 13 HIV-1 proteins. For the interactions of some of the HIV-1 proteins (nucleocapsid, integrase, protease, p1 and env gp160), no opinions were provided. Below, expert opinions collected and the results of applying the maximum likelihood estimator are summarized.

7.4.5 Refining the Literature Curated HIV-1, Human Interactome

In order to estimate the accuracies of experts in labeling, only interactions that were multiply annotated are considered. An expert's labeling accuracy were estimated only if the expert provided labels on the pairs that also received labels from at least one other expert, and if the expert also annotated at least three examples for that class. There were 10 experts in each class. Ultimately, those experts whose labeling accuracies were not estimated were assigned the mean of the available expert qualities.

7.5 Results

Below, we describe first the results of the synthetic data experiments, then the results from applying the method to the real-world data literature-curated PPIs between HIV-1 and human proteins.

7.5.1 Synthetic Data Experiments

The results of two example synthetic data experiments are shown in Table 7.2. There were three experts with various levels of labeling quality; these qualities are listed in in the third column of Table 7.2. For example, the first expert's labeling quality for the 'not direct interaction' class is 0.5, meaning this expert on average labels half of the 'not direct interaction' pairs with a 'direct interaction' label. The probability of an expert labeling an example was set to 0.7 and was uniform across all pairs. The estimations

			global annotation	subset annotation
Class	Expert	True θ	$\hat{\theta}^{\text{mle}}$	$\hat{\theta}^{\text{mle}}$
Not interaction	1	0.5	0.4920 (0.020)	0.4446 (0.087)
	2	0.6	0.6060 (0.105)	0.5945 (0.097)
	3	0.9	0.9069 (0.036)	0.9248 (0.046)
Interaction	1	0.9	0.8768 (0.015)	0.8582 (0.058)
	2	0.7	0.7178 (0.089)	0.7060 (0.035)
	3	0.4	0.3877 (0.032)	0.4446 (0.037)
$<$ MSE $>$ Theta (std):			0.0022 (0.002)	0.0052 (0.039)
$<$ MSE $>$ Uncertainty (std):			0.0023 (0.002)	0.0046 (0.035)

Table 7.2: Two example synthetic data experiments with three experts for global annotation and subset annotation scenarios. In each case, true labels were generated at random, $\mathbf{P}(y = \text{direct interaction}) = 0.5$, and the experiments were repeated n=300 times. In the case of global annotation scenario, each instance was labeled by all three experts, while in the subset annotation case experts decide to label each instance randomly, with the probability of labeling set to 0.7 for each instance. True labeling accuracies (θ) of the experts are given in labeling examples of 'Direct interaction' and 'Not direct interaction' classes, together with the sample mean of estimated labeling accuracies, $\hat{\theta}^{\text{mle}}$, and the standard deviations (std). The average mean squared error in estimating the theta and the uncertainty of the examples are listed in the last two rows.

were better in the global annotation case as expected. $\hat{\theta}^{\text{mle}}$, is $0.0022(\pm 0.002)$ in the global annotation case; this error rate is doubled when subset annotation is applied: $0.0052(\pm 0.039)$. The MSE in estimating the uncertainties were also calculated (see 7.4 for details). These errors were also small, namely $0.0023(\pm 0.002)$ and $0.0046(\pm 0.0035)$ for the global annotation case, where all experts provide a label for each of the interactions, and a subset annotation case, where experts label a subset of interactions.

In order to understand the method's robustness for the number of examples and experts present, the error rates were measured as a function of the number of experts and number of pairs annotated. Figure 7.2 A displays the results of these experiments for the global annotation case when estimating the experts' labeling accuracies. Not surprisingly, the error decreases as more experts are included and more data are provided. Nevertheless, the average MSE of expert labeling accuracies, θ, is $0.0087(\pm 0.0121)$, even in the case in which there are only 3 experts and $n = 100$ data points. Similarly, the error in estimating the uncertainties of the data points is not more than 0.010 (see Figure 7.2 B). Comparison of the error curves for different N reveals the gain in accuracy decreases

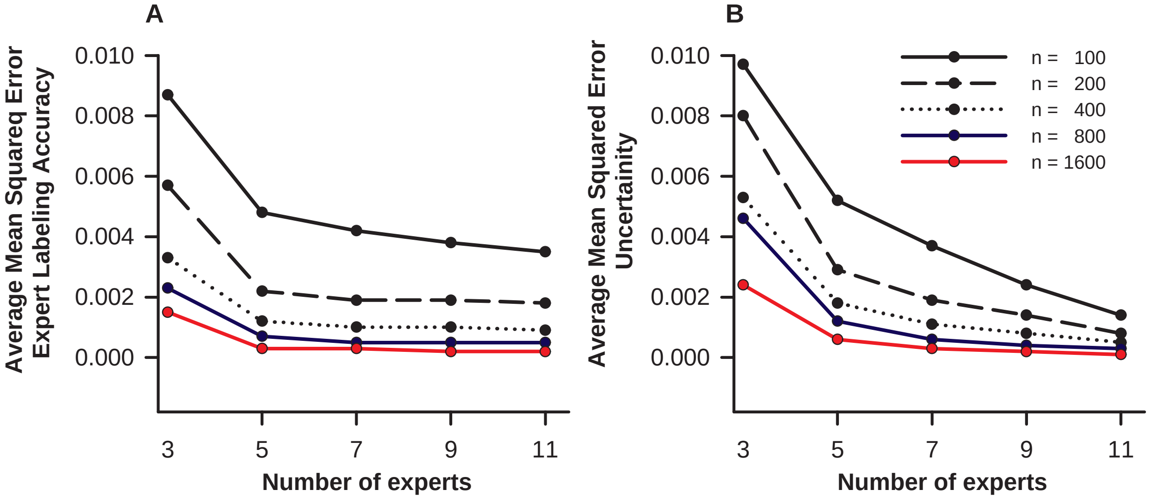


Figure 7.2: The average mean squared errors in estimating a) expert labeling qualities and b) uncertainties is plotted as a function of the number of experts for different numbers of pairs to be annotated. As more data points and more experts are involved in the estimation, the estimator performs more accurately.

in different data regimes. For example, the error in estimating θ decreases by an amount of 0.003 when N is increased from 100 to 200 and there are 3 experts. This difference is only 0.0008 for cases $n = 800$ and $n = 1600$. A similar trend holds for the number of experts; the largest gain in accuracy is observed when the number of experts increases from 3 to 5. The estimation of uncertainties also follow a similar trend (see Figure 7.2 B). Figure 7.3 compares the error rate of labeling accuracy between the global annotation and subset annotation case for the case $n = 800$.

To assess how well the method retrieves the true direct interactions, the precision and recall rates (sensitivity) of the estimator were calculated (see Section 7.4.3.3). The precision and recall of the MLE estimator were compared to four other estimators (described in Section 7.4.3.2). Figure 7.4 displays the precision and recall for different numbers of labelers for the experiments described for Figure 7.2. As can be seen in the figure, single-voting would cover the largest quantity of the true interactions correctly, but would also consider many non-interactions as interactions, therefore displaying a low precision and high recall rate. A similar observation is valid for double-voting. In both cases the pre-

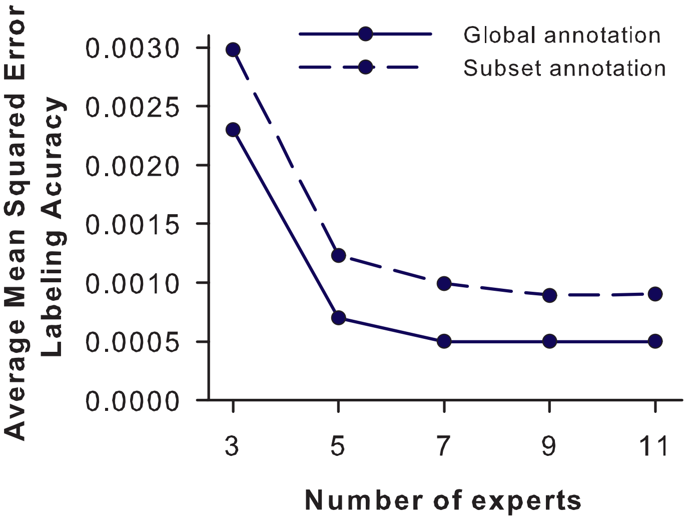


Figure 7.3: The average mean squared errors in estimating expert labeling qualities in subset and global annotation cases for N=800.

cision gets even worse as the number of labelers increase, since the probability of any of the two labelers giving an incorrect label increases. The opposite is true for the all-voting case; the precision is high since the criterion to label a pair as direct interaction is very strict: an agreement between all experts is sought. However, this estimator suffers from low recall. In summary, the all-voting strategy results in high confidence sets, but disregards a large portion of the available data; whereas single-voting or double-voting lead to sets with high coverage but both suffer from high false positive rates. The majority-voting method is a robust one; both the precision and recall rates are high, and additionally, as the number of experts increases, the performance does too. Nevertheless, the maximum likelihood estimator is the best for both precision and recall rates for all numbers of experts. This is probably because the noise is taken into account in our probabilistic framework.

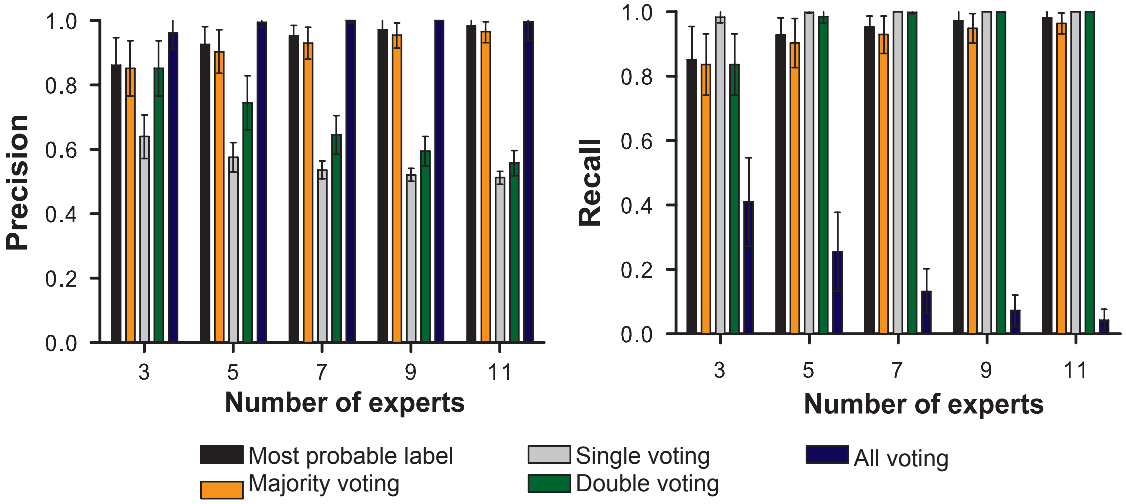


Figure 7.4: Precision and recall rates of different labeling strategies. For description of voting strategies see Section 7.4.3.2.

7.5.2 Refining the Literature Curated HIV-1,Human Protein Interactome

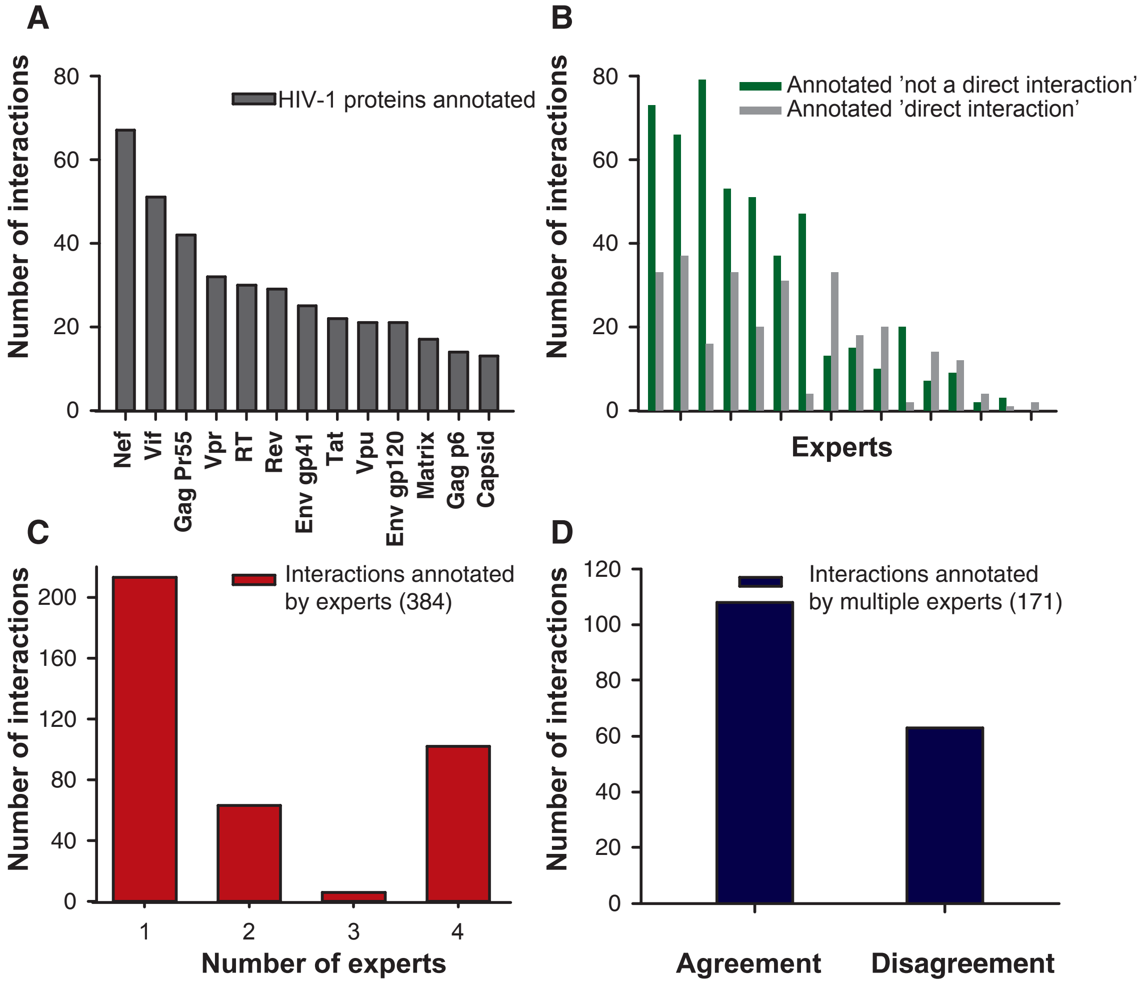


Figure 7.5: Descriptive statistics of expert annotated HIV-1, human protein interactions. a) Number of interactions annotated for each HIV-1 protein b) the number of interactions annotated as 'direct' (green) and 'not a direct interaction' (gray) by each 16 HIV-1 experts c) distribution of number of experts annotating each interaction. Majority of the interactions are annotated by single expert d) distribution of multiple expert annotated interactions in terms of agreement among the experts.

For the various HIV-1 proteins, different numbers of interactions were annotated; nef had 67 interactions annotated, whereas capsid had only 13 (see Figure 7.5). The number of interactions that each HIV-1 expert annotated also varied (see Figure 7.5). The majority of the interactions received only one expert opinion (213/384), whereas for the rest of the interactions (171/384), multiple experts commented on each as seen in Figure 7.5. In cases where an interaction received multiple opinions from different experts, disagreements among the experts was observed. Of all interactions on which multiple experts provided labels, on 37% of them (63/171), experts disagreed on the label type (see Figure 7.5 d). Of all the expert annotated interactions, 299 of them were described by 'interacts' with and/or 'binds' keywords; those have the most potential to be direct interactions. However, strikingly, at least two experts with no disagreement annotated 73 of them as 'not direction interaction'. These results highlight the necessity of reviewing the published interactions with community opinion.

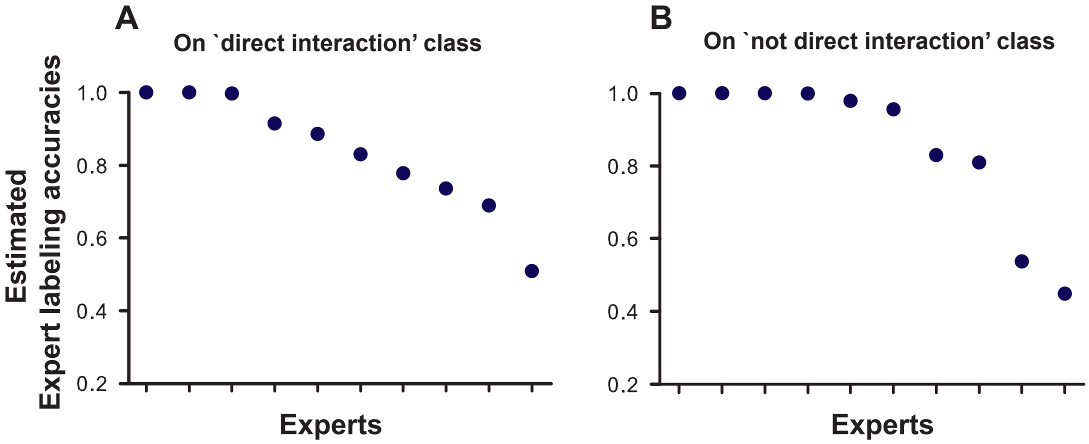


Figure 7.6: The estimated HIV-1 experts labeling accuracies on annotating the protein interactions for a) 'direct interaction' class and b) 'not direct interaction' class.

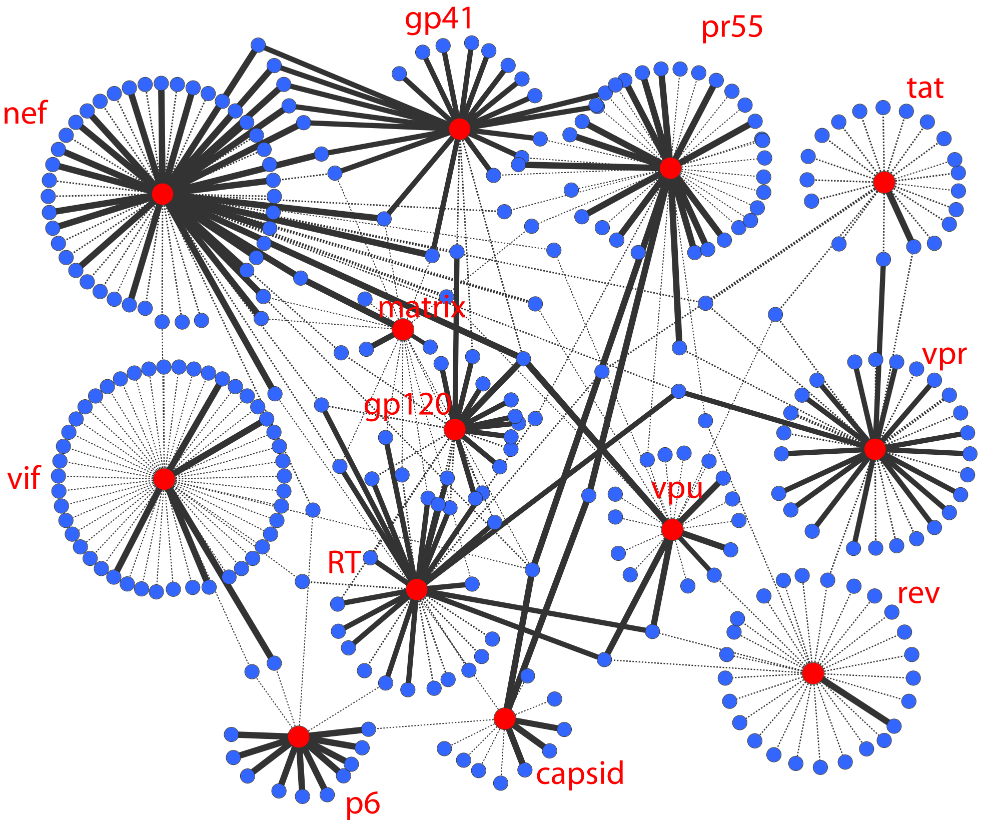


Figure 7.7: Refined HIV-1,human protein interaction network based on the set of HIV-1 expert opinions. Nodes indicate HIV-1 proteins (red) or human proteins (blue); an edge indicates there is at least one publication reporting this interaction and at least one expert gave an opinion about it. The thicker the edge, the higher the probability of the pair being a direct interaction according to the estimates. The solid lines are the interactions where $\mathbf{P}\,(y = \text{direct interaction}) > 0.5$, whereas dashed lines indicate where this probability is < 0.5. The HIV-1 protein's names are placed next to its network of interactions.

Using the interactions that received more than one expert opinion (171 interaction pairs), the experts' labeling accuracies were assessed. There were 16 experts in total, but not all experts provided data for both label types. The estimated label accuracies for the experts are plotted in Figure 7.6. 7 of the 10 experts had a labeling accuracy of more than 75 % accuracy on the 'direct interaction' class; 8/10 had this level of accuracy for the 'not direct interaction' class. Using the expert labeling accuracies, the most probable class label was calculated for all the annotated interactions. For 147 (out of 384) of the reported interactions, there is enough evidence to conclude that they have a direct interaction. Figure 7.7 displays the resulting network. The thicker the edge between two nodes, the higher the probability that it is considered a direct interaction by a given set of experts. The dashed lines indicate those pairs that have a probability of less than 0.5; in other words, there is not enough evidence supporting their direct interaction label according to the experts.

7.5.3 Model Trained with Expert Labels

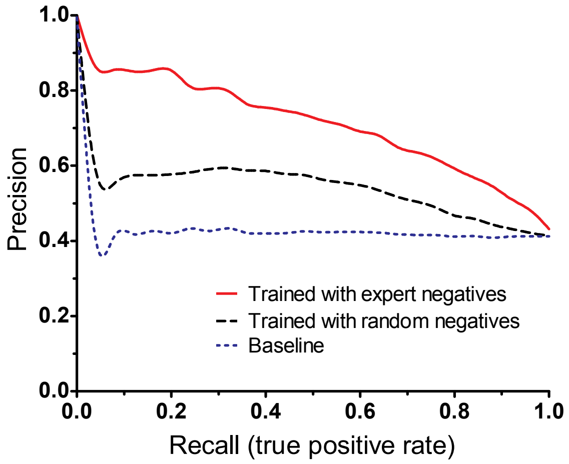


Figure 7.8: Precision recall curves of three different models: i) Model trained with expert labeled positive and negative examples ii) Model trained with expert labeled positive and randomly selected negative examples iii) Model trained with expert labeled positive and negative examples, where the labels of the examples were shuffled. All three models were trained with 42 features described in Chapter 6 and tested on expert labeled positive and negative examples through 3-fold cross validation. The cross-validation experiments were repeated 10 times. The values reflect the averages on the 30 runs.

The estimated labels obtained and analyzed in this chapter provide a high quality set of interaction labels, which include 158 positive examples and 226 negative examples. These labels obtained are especially valuable in training and testing the supervised models. In

	MAP	**stderr MAP**	**AUC**	**stderr AUC**
Trained with expert negatives	0.7144	0.0083	0.7818	0.0052
Expert labeled positive + random negative	0.5466	0.0077	0.6392	0.0084
Baseline	0.4298	0.0075	0.4966	0.0098

Table 7.3: Averages (avg) and standard error (stderr) of MAP and AUC scores over 10 repeated 3-fold cross-validation experiments. Precision recall curves of three different models: i) Model trained with expert-labeled positive and negative examples ii) Model trained with expert-labeled positive and randomly selected negative examples iii) Model trained with expert-labeled positive and negative examples, in which the labels of the examples were shuffled. All three models utilizes the 42 features described in Chapter 6 and they were tested on expert-labeled positive and negative examples through 3-fold cross validation. The cross-validation experiments are repeated 10 times. The values reflect the averages over the 30 runs.

order to judge this, a Random Forest classifier was trained using the expert labels and the 42 features described in Chapter 6. This model achieved 71% MAP score, which is significantly better compared to a baseline classifier, which received 43% MAP score (see Table 7.3). The baseline classifier was trained on the same set of examples but the labels of the training examples were randomly shuffled.

The negatively labeled data are especially valuable as we do not need to resort to randomly selecting the negative labels. As the randomly selected negative labels are likely to be far away from the class boundary, they are easier to classify and more likely to give optimistic estimates of prediction success; the decision boundary learned from them might be far away from the real decision boundary. Conversely, the expert negative labels are also more likely to be in the class boundary since they are examples of functional associations or indirect interactions. Therefore, these negative examples will define a better decision boundary.

In order to judge the negatively labeled expert data's contribution to the model, a third Random Forest classifier was trained. This model uses the positive expert labeled data. However, instead of the expert labeled negative examples a randomly selected set of negative pairs not reported in the NIAID database was used (see Section 5.2.3). Both models included the same number of positive and negative examples. This model performed better than the baseline, 0.55 (± 0.0077) (compare to 43%), but it performed worse than the first model which used the expert labeled negative examples, (compare to 71% MAP). The AUC values are also ranked similarly; the first model achieves an AUC score of 78% (± 0.0052), while the second model only reaches 64% and the baseline is 50% (± 0.0098) (Table 7.3). The precision/recall curves are well separated for all the three models, where the model trained with expert labeled negative and positive data outperforms the other two models (see Figure 7.8). Note that the models in Chapter 5 and Chapter 6 were tested with data where the class distribution was skewed. The negative to positive example ratio was 100:1. Here, the expert-labeled data contains positive:negative labels in a positive:negative labels in a 1.5:1 ratio; therefore, the AUC and MAP scores are higher and the precision/recall curves are elevated in this experiment.

All three models were tested on expert labeled positive and negative training examples in a 3-fold cross validation setting. The cross-validations were repeated 10 times, where at each repeated run, the splitting of the data is different and random. These empirical results strongly indicate that the classifier benefits from the expert labeled data.

7.6 Conclusions

In this chapter, a gold standard dataset of HIV-1, host interactions was presented where the confidence of each interaction was estimated probabilistically. To arrive these confidence scores, I developed a maximum likelihood approach to estimate the experts' labeling accuracies in an unsupervised setting. This approach is general and can be used in other crowd sourcing applications, in which noisy labels from multiple experts are available and there is no benchmark data to estimate labeler qualities. Using the maximum likelihood estimate of the expert labeling qualities, I calculated the probability of being a direct interaction given a set of expert opinions for each interaction. These labels obtained are valuable in training and testing the supervised models. A Random Forest model trained with using the expert labels performed with a MAP score of 71%. This is significantly better than a baseline classifier, which was trained using the same training set but shuffled class labels. This model performed with a MAP score of 43%. Especially the negative labels are valuable. By training a model using the expert labels and randomly selected negative labels, a score of 64% was obtained, which was significantly smaller than 71% in the case where expert labeled negative examples were used.

Chapter 8

Multi-Task Learning for Predicting Host,Virus Interactions

8.1 Overview

In the supervised models described in Chapters 5, 6 and 7, the task of predicting host-virus interactions was cast as a single-prediction task. In this formulation, the PPIs of all viral proteins were pooled together and a single model was learned to define the whole host-virus interactome. However, as the viral proteins undertake different functions in the viral replication cycle, they might well be drawn from different distributions; and pooling the training data together might disregard these differences. An alternative would be to learn different models for each viral protein. However, small sample size for each of the viral proteins' in the expert labeled interaction data impedes the construction of reliable separate models. In order to overcome the data scarcity issue while considering possible differences in data distribution across viral proteins, we developed a multi-tasking learning strategy. In this model, a learning task was defined for each of the viral proteins separately, but these tasks shared their training data proportional to task relatedness. Herein, the relatedness of the tasks were defined based on the viral proteins' role in the replication cycle. A multi-task Random Forest learning method is presented, which modifies the regular Random Forest. In the regular Random Forest classifier, several bootstrap samples were created from the input training examples. In the multi-task Random Forest classifier, for a given task, the examples are drawn from the pool of all examples with the probability proportional to their task relatedness to the other tasks. This modification lead to more accurate predictions for 8 of the 10 HIV-1 viral proteins as compared to those derived from the single model learned for the HIV-1, host proteins.

8.2 Methods

8.2.1 Data

HIV-1 protein	**Number of Expert Labeled HIV-1,Host PPIs**	
	Negatively Labeled	Positively Labeled
Env gp120	10	11
Env gp41	3	22
Gag capsid	8	5
Gag matrix*	15	2
Gag p6	4	10
Gag pr55	22	20
Gag nef	35	32
Reverse transcriptase	10	20
Rev*	28	1
Tat*	20	2
Vif	46	5
Vpr	13	19
Vpu	12	9
Total	158	256

Table 8.1: 13 HIV-1 proteins had at least one label estimated from expert opinions (see Chapter 7). Table lists the number of positively and negatively labeled examples. The star marks the proteins for which a classifier is not trained due to limited number of positive examples.

In this section, the HIV-1, host PPI data labels obtained through expert annotations were used because these are high-quality labels. The details of collecting expert opinions and estimating the final labels probabilistically are provided in Chapter 7. This data set included PPIs between host proteins and 13 HIV-1 proteins. The number of positively and negatively labeled examples for each viral protein is given in Table 8.1. On 3 of the 13 viral proteins namely, tat, matrix and rev, there were less than three PPIs positively labeled examples, which hindered an evaluation through 3-fold validation. Therefore, no models were trained for these three HIV-1 proteins. Nevertheless, the labeled data for them still contributed to the training of the models for the other viral proteins.

8.2.2 Multi-Task Random Forest Classifier

The problem of learning host, virus interactions is cast as a multi-task learning framework. In this framework, each task is to learn a model specific to each viral protein considered. In this setting, tasks share training examples with each other. Let L be the set of training examples including all viral proteins' interactions, and N be size of L. Let V be the number of viral proteins that have some training data in L and $K < V$ be the number of tasks to be learned. $\forall$ tasks $k \in \{1..K\}$, a multi-task Random Forest classifier, f_k, is learned. A multi-task Random Forest classifier differs from the regular Random Forest Classifier in the way the training examples are drawn in the learning phase. The Random Forest classifier is an ensemble learner, which learns several decision trees using bootstrap samples of the input training data (see Section 5.2.2.2 Algorithm 1). In the multi-task learning, this bootstrapping step is changed as follows.

First, based on biological knowledge, relatedness of the viral proteins are defined. It was based on their pairwise functional relatedness, but other definitions can be adopted. $w_{i,j}$ denotes the similarity of the viral protein i to viral protein j and is in the [0-1] range. The probability of viral proteins of type i being drawn in learning the task for viral protein j is proportional to $w_{i,j}$. Formally, let $A_{i,j}$ be the probability that of a given interaction pair of viral protein i being drawn in the bootstrap sample of the viral protein j's task. This probability is:

$$\mathbf{P}\left(A_{i,j}\right) = \frac{w^{\alpha}_{i,j}}{\sum_{v \in \{1..V\}} w^{\alpha}_{v,j}} \tag{8.1}$$

where α is a scaling parameter and is in the range [0,1]. As α decreases, the probability distribution gets closer to the uniform distribution and all examples are drawn with equal probability. The algorithm is summarized in Algorithm 3.

Algorithm 3 Random Forest Classifier For Multi-task Learning

```
N be the number of examples in the training data, and L be the set of training data,
d the number of features, B the number of trees in the forest, $n_{min}$ minimum number
of examples allowed on a node, m number of features to be used for determining the
splitting feature. Let K be the number prediction tasks. Let $\{1..V\}$ be the set of viral
proteins. Let w be the similarity matrix that defines similarities between viral proteins.
Similarities range in [0,1] and $\alpha$ is a scaling parameter and assumed to be given.
1. For each task k grow a Random Forest:
for $b = 1$ to B do
   2. Construct a Random Forest tree for task k, $T_{k,b}$:
   a) Draw a bootstrap sample of $Z^*_{k,b}$ of size N from the training data such that for a
   $i, j \in V$, $\mathbf{P}\left(A_{i,j}\right) = \frac{(w_{i,j})^{\alpha}}{\sum_{v \in \{1..V\}} w^{\alpha}_{v,j}}$
   b) Grow a tree $T_{k,b}$ using $Z^*_{k,b}$. In growing the tree at each terminal node of the tree
   recursively apply the following steps:
   repeat
      i. Select $m \leq d$ features among the d features at random as candidates for splitting.
      ii. Pick the best splitting feature among the m features based on Gini impurity
      index.
      iii. Split the node based on the chosen feature into two.
   until $n_{min}$ is reached
end for
2. Random Forest for task k is the ensemble of trees $\{T_{k,b}{}^{B}\}$
3. Let $\hat{y}_{k,b}$ be the class prediction of the $b^{\text{th}}$ tree in the Random Forest of task k, then
the label for a new example x will be:
$\hat{y}^B_k$ = majority vote$\{(\hat{y}_{k,b}(x)\}^B$
```

The similarity of the viral proteins was defined by the relatedness of their functions using prior biological knowledge. I defined the similarity by considering whether two proteins take part in the same step of the replication cycle. In Figure 8.1 the cells are

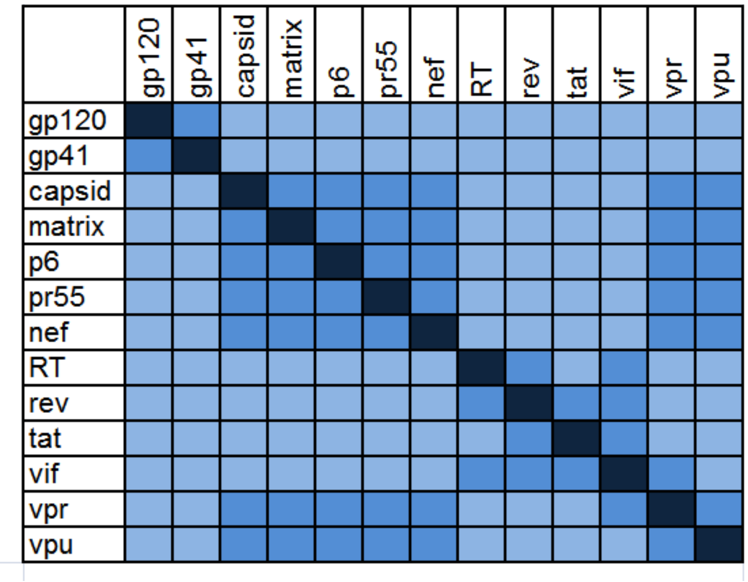


Figure 8.1: Heatmap showing the similarities between viral protein pairs. The similarity is based on the functional relatedness of the viral proteins. The darker the color, the higher the similarity; specifically, the lightest blue cells are valued at 0.1, the medium blue cells at 0.5 and the darkest color cells at 1.

color coded based on these similarities. The very light blue color indicates a similarity of 0.1 and a medium color indicates 0.5 and a dark blue color indicates similarity of 1. The similarity matrices were selected based on the mean average precision (MAP) score of the validation set. A thorough parameter search can be conducted and embedded in the learning phase, if there is significant training data.

8.2.3 Features

The final feature set, which was described in Sections 5.2.4 and 6.2.2 and evaluated in Chapter 6 is utilized. This set includes 42 features derived from various proteomic and genomic data sources.

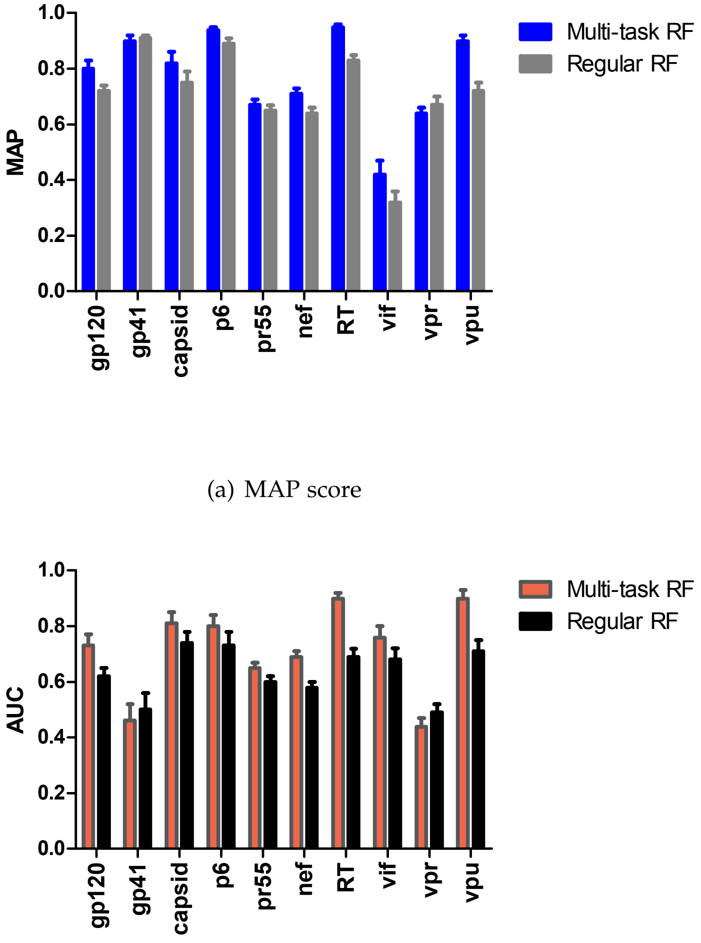
8.2.4 Evaluation

The two metrics, AUC score and the MAP score, were employed to evaluate the classifier. Each task was evaluated separately. The empirical evaluation is conducted in a 3-fold cross-validation setting. The cross-validation experiments were repeated 10 times; in each case the data were split randomly. The results were averaged over these 30 runs. The negative to positive data ratio was 1.4:1, which is reflected in the scores. Therefore it is not possible to directly compare these scores to the MAP and AUC scores obtained in other chapters, where the test data were skewed by a different ratio.

8.3 Results

8.3.1 Performance of the Multi-Task Random Forest Classifier

For 13 of the 17 HIV-1 proteins we had expert-labeled positive and negative data (Table 8.1). Three of these proteins (tat, rev, matrix) had less than 3 positive examples; therefore, as their performances could not be measured no models were learned for them. However, these proteins contributed to the training phase of the other 10 proteins, which had sufficient expert-labeled data. Similarity among proteins was defined for the viral proteins based on the relatedness of the functions they perform. I trained and tested the multi-task Random Forest classifier and compared it to the case where a single model was learned by pooling all the viral proteins' interactions together. Both models were separately tested for each viral protein on the same testing examples separately for each viral protein. Figure 8.2 shows the MAP and average AUC scores for the 10 proteins. For 8 of the 10 proteins except env gp41 and vpr, the multi-task learner outperformed the model where all interactions were learned globally.



(a) MAP score

(b) AUC score

Figure 8.2: Performance of the multi-task Random Forest model compared to the case when a single model is learned. MAP and AUC scores are averages over 10 repeated 3-fold cross validations.

8.3.2 Model Trained with Multi-Task Random Forest Classifier

We trained a final model using all the available training data for the 10, HIV-1 proteins. The network with a Random Score cutoff of >0.5 included 3797 HIV-1, virus interactions involving 2719 unique human proteins.

One of the interesting predictions was the interaction of gag p6 with the host protein ATPase VPS4. The model strongly predicts that the two proteins interact (third highest ranked prediction). There was no HIV-1 interaction reported in the NIAID database that involves Vps4A. Yet, one of the genome-wide RNAi screens identified Vps4A as the host factor. Moreover, Urata et al. [199] showed that the enzymatic activities of Vps4A and Vps4B are required for efficient budding of HTLV-1 and gag p6 is known to be involved in viral budding strengthening the hypothesis for this prediction to be true.

8.3.3 Gene Ontology Term Enrichment Analysis

I assessed the enrichment of GO terms in the set of predicted interaction partners. Partial lists of significantly enriched GO terms for molecular process, function and cellular location are given in Tables 8.2, 8.3, 8.4 respectively. Many of the highly enriched molecular processes were related to transportation. This is different to the previous models' (see Chapter 5 and Chapter 6) enriched terms. In the molecular function case, binding , kinases and nucleotide binding were among the top enriched functions.

GO term ID	**GO term name**	**p-value**
GO:5515	protein binding	2.75e-33
GO:5488	binding	3.26e-20
GO:4674	protein serine/threonine kinase activity	2.01e-13
GO:4672	protein kinase activity	3.15e-13
GO:16773	phosphotransferase activity, alcohol group as acceptor	9.39e-13
GO:16301	kinase activity	2.37e-11
GO:32555	purine ribonucleotide binding	7.47e-11
GO:32553	ribonucleotide binding	7.47e-11
GO:17076	purine nucleotide binding	2.88e-10
GO:16772	transferase activity, transferring phosphorus-containing groups	3.05e-10
GO:166	nucleotide binding	5.83e-10

Table 8.2: Enriched GO biological function in the set of predicted human interaction partners.

GO term ID	**GO term name**	**p-value**
GO:3034	intracellular signaling pathway	1.99e-20
GO:15031	protein transport	8.39e-19
GO:45184	establishment of protein localization	1.89e-18
GO:8104	protein localization	4.74e-16
GO:33036	macromolecule localization	2.96e-15
GO:16192	vesicle-mediated transport	3.23e-15
GO:61024	membrane organization	1.80e-12
GO:35556	intracellular signal transduction	1.91e-12
GO:16044	cellular membrane organization	2.87e-12
GO:46907	intracellular transport	1.56e-10

Table 8.3: Enriched GO molecular processes in the unique set of human proteins that involve in the predicted viral, host protein pairs.

GO term ID	GO term name	p-value
GO:737	cytoplasm	1.23e-39
GO:5622	intracellular	1.86e-32
GO:44424	intracellular part	6.78e-26
GO:5829	cytosol	3.20e-21
GO:30117	membrane coat	2.94e-16
GO:48475	coated membrane	2.94e-16
GO:44444	cytoplasmic part	1.20e-15
GO:30118	clathrin coat	2.26e-13
GO:43227	membrane-bounded	organelle 5.83e-13
GO:43231	intracellular membrane-bounded organelle	5.83e-13
GO:10008	endosome membrane	9.58e-13

Table 8.4: Enriched GO cellular components in the set of predicted human interaction partners.

8.4 Conclusion

In this section, a multi-task learning framework was presented for predicting viral host interaction partners. The feature set described in Chapter 6 was utilized and training data included PPIs whose labels were estimated from the expert opinions (see Chapter 7). Each viral protein's interaction was modeled as a separate task and each task shared across training examples based on their relatedness. This is achieved by a simple modification in the bootstrapping step of the Random Forest classifier. The empirical evaluation showed that for most of the viral proteins, using this multi-task framework improves the model performance. Here we defined the relatedness of the tasks based on the functional similarity of the viral proteins. Alternative criteria can be employed, or this matrix can be learned simultaneously if there is sufficient data to explore the parameter space.

Chapter 9

Conclusions and Future Work

Numerous human diseases are caused by viral infections. Our lack of understanding of the intimate relation between the virus and its host makes the development of therapies difficult. Protein-protein interactions are key players in every cell function, both within and between organisms, at every level of cellular function. Comprehensively identifying these interactions enables us to detail how cellular processes take place. Past experimental and computational research largely focused on identifying interactions within single organisms. Computational approaches, first in model organisms, and later in human, have helped experimental efforts in revealing parts of the protein interactomes. On the other hand, characterizing the interspecies interactomes on a system wide level has only been a recent focus. High-throughput experimental techniques are being adapted to handle the interactions of both organisms at the same time. However, there is still no single cost-effective and highly accurate experimental technique to identify interactions on a large scale. As was the case for intra-species protein interactomes, computational methods could be utilized to accelerate experimental endeavors. My work presented in this thesis aims to accelerate efforts to identify interspecies interactomes, and is one of the early studies of the inter-species prediction task. In this section, I will provide an overview of the thesis. Overall, this chapter summarizes the work conducted in this thesis, and points out potential directions for future studies.

9.1 Thesis Summary

Working towards defining the host-virus interaction network, this thesis focused on predicting host-virus interactions. Specifically, I concentrated on the HIV-1-human interactome because it is clinically important and represents the system with the richest experimental data available; however, the methods and data curation techniques presented can easily be extended to other inter-species systems as pertinent data become available. Throughout the thesis, I employed a machine learning perspective. The task of predicting PPIs were formulated in a binary classification framework, where each possible protein pair falls into one of two classes, the 'interacting protein pairs' (positive class) and the 'non-interacting protein pairs' (negative class). The classifiers were learned in a supervised setting. In developing these predictors, several data and methodology related challenges were handled. Specifically, I identified and compiled biological sources that can be utilized as predictive features, collected expert annotations to create gold-standard PPI labels, devised a computational method to aggregate noisy expert labels and presented a novel multi-task learning framework. The models were iteratively refined by improving the labeled data, features and computational models.

9.1.1 First Supervised Model for Predicting HIV-1, Host PPIs

Chapter 5 describes the first supervised model. A Random Forest classifier was employed to learn to distinguish interacting proteins from non-interacting pairs. One challenge to building such a system is identifying biological information that can serve as predictive features. Several data types that have been useful in the intra-species prediction task (see Section 3.1) are not directly applicable to the host-virus setting. Therefore, identifying information that is predictive in distinguishing interacting protein pairs from non-interacting ones is important. Going through an extensive curation process, experimental results pertinent to host-virus interactions were identified and retrieved from databases or published articles. These biological data included a wide array of information such as gene expression profiles of HIV-1 proteins during infection, sequence-motif pairs found frequently in interacting proteins, functional similarities, etc. In encoding such biological information as features, I also took into account the known interactions among the host proteins. For example, as the viral proteins might be similar to the host proteins' interaction partners, I included viral proteins' similarities to the host proteins. These similarities were based on sequence, translational modifications, function, molecular process and cellular location similarities. My results, as described in Chapter 5 and Chapter 6, demonstrated that these features, which take into account the cellular contexts of the host, are especially informative. For instance, network node properties of the human proteins in the human PPI network were among the most predictive features. Chapter 5 described this biological information and how it was encoded as features. A Random Forest classifier was learned using this feature set and the set of labels. The model was evaluated using cross-validation and achieved a MAP score of 23%, much better than that of the random baseline models. The predictions were also compared to external biological information, including genome-wide RNAi screens that were available at the time and the set of proteins detected in the budding virion. 21 host proteins were tested as to whether they colocalize with vpr and capsid, using single live cell imaging techniques. The colocalization experiments provided evidence on the validity of the predictions, highlighting that the computational methods could provide experimentally testable hypotheses.

9.1.2 Improved Prediction Model for HIV-1, Host PPIs

In Chapter 6, I improved upon the first supervised model described in Chapter 5. This model made use of the same computational setting and data labels as described in Chapter 5. The improvement was achieved by incorporating new biological information that became available. One of the information sources was the four genome-wide RNAi screens results, which identified sets of host factors that are required for virus replication. Also, a large-scale affinity purification mass spectrometry experiment result was conducted with HIV-1 proteins to identify host interactions. Sets of human proteins detected in budding virions and those that interact with other viruses were included. This information was encoded in the system by taking into account cellular interactions, known pathways and complexes. The resulting model improved upon the first model by a 12% increase in the MAP score. Of the newly incorporated features, those utilize the RNAi screen results were the most predictive. The resulting predictions included proteins that were independently experimentally validated by others.

9.1.3 Obtaining High-Quality Interaction Labels

A challenge to developing robust PPI networks is obtaining high-quality positive and negative labels. The supervised models described in Chapter 5 and Chapter 6 made use of a subset of HIV-1, human protein interactions reported in literature as positive examples. The negative sets were obtained by randomly pairing viral, host protein pairs from the set of all possible pairs not reported in the literature. Neither of these approaches is ideal for creating high-quality positive and negative datasets. Discussion with HIV-1 experts revealed that the interactions provided in the NIAID HIV-1, human protein interaction database [134, 135] are not considered reliable. When I analyzed the number of supporting publications for the PPIs reported in the database, I found that 44% of all of the PPIs are reported only in a single publication (see Chapter 7). The lack of follow-up studies hints at the possibility that for many of these interactions, there may not be sufficient experimental evidence to support their direct interaction.

In Chapter 7, I addressed the label quality issue and obtained a high-quality set of positive and negative labels of HIV-1, human direct PPIs. Specifically, I i) collected opinions of HIV-1 experts about the interactions reported in the literature and ii) formulated a probabilistic framework to assign reliability scores to interactions based on noisy, subjective expert opinions. Assessing the data quality of PPIs from small-scale experiments requires a complex judgment about the methods and results of each specific study. Some experimental techniques more conclusively identify functional relations, while others more conclusively identify direct interactions; techniques do not work uniformly well across all proteins. In addition to the variability of the powers and limitations of each technique, the condition under which a study is conducted is important. Such a complex judgment may be provided best by domain experts who would take into account all of these parameters. In order to arrive at a high-quality label set for the literature curated HIV-1, human PPIs, a crowdsourcing approach was taken. HIV-1 experts were presented with the accumulated published evidence and asked to annotate interacting pairs with labels based on whether the interaction is supported by enough evidence to conclude that the pair represents a direct physical PPI. In cases where a PPI received opinions from multiple experts, disagreements among experts were common. This is understandable as experts have different biases and expertise. Therefore, the challenge is how to arrive at an accurate estimate of the label type, given noisy, subjective expert opinions. A maximum likelihood approach was taken to estimate the experts' labeling accuracies for each label type. Next, these estimated labeler accuracies were used to calculate the probability that the interaction is a true direct interaction. In this model, annotators were not required to have the same labeling quality; moreover, it was allowed that experts may have different labeling qualities for the label types 'interacting' and 'non-interacting'. The computational model described in Chapter 7 is not limited to curated data for HIV-1 protein interactions, but is applicable to cases where multiple noisy labels need to be aggregated, a common setting for crowdsourcing applications. In this chapter, our results also showed how the expert curated data improved the supervised learning model. The negatively labeled example set especially contributes to model performance and outperforms the model that used the randomly selected negative examples significantly.

9.1.4 Multi-Task Learning for Host,Virus PPI Prediction

In the above classifiers, all the viral proteins' interaction data were pooled together and solved the problem as a single task. However, the viral proteins undertake different functions and participate in different parts of the replication cycle, which implies they can be treated as drawn from different distributions. This necessitates building different models for each viral protein. However, the lack of sufficient data for many of the HIV-1 proteins impedes the construction of separate models for each protein. In order to overcome the data scarcity issue while not disregarding possible differences in data distribution across viral proteins, I developed a multi-task learning strategy. In this model, single tasks (learning the protein-protein interactions of each viral protein) are learned, but these tasks shared training examples proportional to their relatedness. This multi-task Random Forest model represents a modification in the training phase of the regular Random Forest model. In the Random Forest classifier, the training examples are drawn randomly with replacement to create bootstrap samples when building the decision trees. In the multi-task version, the training examples are drawn from a modified distribution where the probability of each example being drawn is proportional to the examples' relatedness to the viral protein at hand. Such a multi-task framework leads to more accurate predictions and provides a rich set of hypotheses on the HIV-1, host predictions.

9.2 Practical Usage of Predicted Virus-Host Interactions

The predicted virus-host interactions can be used to guide detailed experimental studies by reducing the hypothesis space of all possible interactions to a tractable set. Predictions can be utilized in different ways depending on the biological interest in and the prior knowledge of the predicted pair. Ultimately, the interactions can be used for drug discovery because new targets and approaches to combat viral infections are urgently needed. Below, suggested areas in which the predictions can be useful are discussed.

9.2.1 Novel Predicted Pairs

Novel predicted pairs are those interactions that have not been previously reported in the literature. Such new pairs are candidates for biological hypotheses on previously unrecognized virus hijack mechanisms. A virologist looking for new antiviral drug targets or seeking to understand viral function and host defense would be most interested in such novel interaction pairs. The functional significance of the predictions can be checked by their effects on viral infectivity, and by more specific assays investigating in-depth mechanisms of viral processes. Subsequently, more detailed functional and structural experiments can be designed to dissect the details of how the uncovered virus-host's protein-protein interaction contributes to the viral replication cycle. The results of these experiments have the potential to bring new insights not only about virus biology but also about the biology of the host. As cellular proteins function in multiple different ways, this may help annotate them with previously unrecognized functions.

9.2.2 Pairs with Prior Experimental Evidence

Structural studies, such as NMR and X-Ray, are powerful techniques for characterizing the bound configurations of the host-virus complexes, but are also extremely labor extensive and time consuming. Therefore, in contrast to virologists, structural biologists are interested in pairs only if there is extensive experimental evidence of the validity of the interaction and its functional relevance. However, in the case of HIV-1, experimental studies and functional screens identified thousands of putative interactions and functional associations, leaving the structural biologist with the challenge of deciding which pair to pursue. The ranking of predicted pairs, including those that have been reported elsewhere with some evidence, can help the structural biologist prioritize the list of interactions. In this way, the reported interactions can be stratified and true, direct physical interactions can be differentiated from indirect interactions.

9.2.3 Utilizing the Information Related to Features

Analysis of the features of predicted pairs can be used to design biological experiments. For instance, if a predicted pair includes sequence motif-domain pairs that have been frequently observed in interacting protein pairs (see ELM-ligand feature in Chapter 5), these sites can be used to design functional experiments. These putative sites can be mutated and the effect of mutations on different parts of the replication cycle can be monitored. Furthermore, active sites for binding can be identified. Based on the knowledge gained regarding the active sites, antiviral drugs can be developed to prevent the binding event or alter binding affinity.

9.2.4 Identifying New Drug Targets

Ultimately, the predicted list and the follow-up experiments aim to identify host factors that can be targeted to prevent viral infection. Current antiviral drugs suffer from drug resistance problems. Through mutations, the viral genome can render an otherwise effective compound ineffective. Antiviral therapies that target host-virus interactions are promising since cellular factors would not be expected to mutate under antiviral drug pressure [13]. The predicted list of interactions, once experimentally validated, can be used to design drug targets to block the host-virus interaction. To prevent side effects, the network of interactions can be used as an additional source of information.

9.2.5 Alternative Pathways

By supplementing the known signaling pathway information data with the predicted interactions, new hypotheses can be generated regarding virus-targeted pathways and the functional consequences of viral pathway interception. This is the input needed to design strategies on how to circumvent this effect. One approach is the so-called "alternate pathway hypothesis" [200]. In this approach, simple paths that start with a protein, such as receptor that does not receive an input from another protein, and end with a protein, such as a transcription factor, are defined. Paths may contain proteins that interact with HIV-1 proteins, potentially disrupting pathways important for cellular signaling. The idea is to find alternative paths between the same start and end points that do not traverse any protein that can interact with an HIV-1 protein, counterbalancing the effects imparted by the HIV-1 interactions on the signal transduction pathways. We supplement the combined interaction and signaling pathway map with functional information, namely which proteins are known drug targets and which proteins have shown an effect on HIV-1 infectivity and other functions upon siRNA silencing. As an example, the cholesterol biosynthesis pathway is not targeted by any of the known interactions, but according to our predictions, the HIV-1 protein tat interacts with farnesyl-diphosphate farnesyltransferase 1 (FDFT1). The pathway also includes a protein that is identified in one of the HIV-1 genome-wide RNAi screens as being functionally important. The cholesterol synthesis pathway would be a good candidate to search for a drug target because it contains alternative paths with known drug targets and has already been functionally linked to HIV-1 biology through the presence of the siRNA gene. Additional evidence supporting such a functional link is given by numerous statistics showing AIDS patients' increased risk for arteriosclerosis. This is an example where an interaction predicted by my model was used in secondary analysis to help generate a new way to identify effective drug targets.

9.2.6 Developing Broad Spectrum Antiviral Drugs

Since current antiviral drugs are designed to target a specific viral enzyme, they have a very narrow spectrum and can only treat specific viral species or subtypes. An in-depth comparison of responses of diverse hosts to the same pathogen, and of the same host to diverse pathogens, should allow the identification of novel avenues for broad-spectrum vaccine development and drug discovery. In this sense, not only host-virus interactions but also transcriptomic data collected upon infection would be useful. For example, if diverse host cells increase expression levels of surface-expressed proteins in response to diverse pathogens, a vaccine raised against this protein should identify these cells as abnormal, priming them for recognition by the immune system. Further, if there are uniform weak points in the host cell response to pathogens or if pathogens target the same proteins and/or pathways, we can potentially use pharmacological means to help the host cell restore at least part of its normal functioning in the presence of pathogens, regardless of the type of pathogen. Such an approach is generally referred to as adjuvants [201].

9.3 Future Work

There are several potential directions for future extensions of this research. Some of them are outlined below.

9.3.1 Predicting Interactions in Other Host-Virus Systems

Along with HIV-1, there are many other clinically important viruses on which computational models could shed light on their interaction with the human host. The binary classification setting I provide and most of the features I derive can be extended to predicting other virus-host PPIs. The limiting step will be the availability of the labeled data and biological pathogen-specific biological information sources to be used as features.

Given that all pathogens encounter similar barriers when using the human cell as a host, they are likely to recruit similar strategies. For instance, enveloped viruses complete their replication cycle through budding from a cellular membrane. Some enveloped viruses are known to make use of the vacuolar protein sorting pathway for budding from the host cell membrane via interaction with the protein in the pathway through similar sequence motifs [202]. Also it has been shown that proteins that are targeted by different proteins are frequently exists in the pathway [182]. Therefore, I hypothesize that knowledge across different viruses can be exploited if they bear similarities. Computationally, one would formulate the prediction of interactions between different pathogens as individual learning tasks. Each of these tasks will be related to each other based on the similarities of the viruses. The multi-task learning framework presented for prediction of PPIs of different viral proteins can be applied. Such a formulation might help to overcome the data scarcity issues that hamper building virus-host models currently.

9.3.2 Predicting Functional Consequences of PPIs

In this thesis, the goal was to predict direct physical PPIs between host and virus proteins. In the biological sense, these interactions can lead to different functional outcomes such as phosphorylating, stable binding, acetylating or degradation. In lieu of predicting the general label of direct physical interaction, one could try to predict a more detailed label type. For this aim, knowledge on functional sequence motifs and domains and annotations can be utilized. Such a richer formulation could provide biologists with more detailed hypotheses to design better experiments.

9.3.3 Crowd-Sourcing for Refining PPI Databases

In Chapter 7, I presented my results on collecting expert opinions on the HIV-1 and host-protein interaction dataset in order to develop high quality interaction data. There are different lines of research related to this in terms of both computational approaches and application to broader PPI databases. In combining subjective expert opinions, I presented a maximum likelihood approach for assessing each expert's labeling quality. Firstly, in this work it was assumed that expert opinions are independent of each other given the class labels. This assumption might not always hold, as collaborators are more likely to share similar opinions; for instance, their biases toward a pull-down assay might be similar. To take into account such relationships, one could model the expert's relationships with each other, possibly by forming a co-authorship network. Given this network structure and multiple expert opinions on the unlabeled interactions, the expert labeling accuracies could then be estimated. This is an interesting problem that could provide diverse applications in other crowd-sourcing settings where the annotators are related.

A second direction would be to extend the 'crowd-sourcing' idea on literature curated interactions to other databases. The lack of confidence scores is a common problem in several other widely used, larger PPI databases. Most of the public PPI databases catalogue and present the experimental details of each study, which in theory allows the user to review the information and use his or her judgment about the level of confidence in the interaction. However, in many systems' biological application, the researcher uses thousands of interactions at once, making it impractical for the investigator to review these interactions. This leaves the database user with the challenge of designing her own heuristics to extract an accurate interaction set based on the type of experiment or the number of publications. Although the number of publications reporting an experiment and type/scale of the experiment are both indicators of how well the data are supported by evidence, there are several other factors that would affect confidence in the validity of an interaction such as the experimental conditions. And experts could weigh these factors differently. In order to arrive at reliability scores, databases could provide an interface in which users provide their labels on how well existing experimental data supports the validity of the protein interaction. The collected data in turn can be used to arrive scores for the PPIs deposited in the database. The computational method presented in Chapter 7 solves this problem. Nevertheless, implementing such a task would present several practical challenges, i.e expert willingness to participate and identification of experts in each field. There could be ways to overcome by crediting database contribution.

Appendix

List of papers published

1. **O. Tastan**, E. Yu, M. Ganapathiraju, A. Aref, A. J. Rader, and J. Klein-Seetharaman. Comparison of stability predictions and simulated unfolding of rhodopsin structures. *Photochem Photobiol*, 83(2):351-62, 2007.

2. **O. Tastan**, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman. Prediction of interactions between HIV-1 and human proteins by information integration. *Pacific Symp Biocomput*, pages 516-27, 2009.

3. **O. Tastan**, J. Klein-Seetharaman, and H. Meirovitch. The effect of loops on the structural organization of alpha-helical membrane proteins. *Biophysic J*, 96(6):2299-312, 2009.

4. S. Balakrishnan, **O. Tastan**, J.G. Carbonell, and J. Klein-Seetharaman. Alternative paths in HIV-1 targeted human signal transduction pathways. *BMC Genomics*, 10 Suppl 3(10):S30, 2009.

5. N. J. Venkatachari, L. A. Walker, **O. Tastan**, T. Le, T. M. Dempsey, Y. Li, N. Yanamala, A. Srinivasan, J. Klein-Seetharaman, R. C. Montelaro, and V. Ayyavoo. Human immunodeficiency virus type 1 vpr: oligomerization is an essential feature for its incorporation into virus particles. *Virol J*, 7:119, 2010.

6. Y. Qi, **O. Tastan**, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics*, 26(18):i645-52, 2010.

7. I. Singh, **O. Tastan**, and J. Klein-Seetharaman. Comparison of virus interactions with human signal transduction pathways. *In Proceedings of the First ACM Int Conf on Bioinformatics and Computational Biology, BCB '10*, pages 17-24, 2010.

8. Z. Zhao, J. Xia, **O. Tastan**, I. Singh, M. Kshirsagar, J. Klein-Seetharaman, and J. G. Carbonell. Virus interactions with human signal transduction pathways. *International Journal of Computational Biology and Drug Design*, 4(1), pages 83-105, 2011.

List of papers in submission

1. **O. Tastan**, J. Carbonell, and J. Klein-Seetharaman. Assessing confidence in curated protein interactions in a probabilistic framework through community opinion.

2. **O. Tastan**, S. Jäeger, J.G. Carbonell, N. Krogan, and J. Klein-Seetharaman. Improved prediction of HIV-1, human protein interactions.

3. **O. Tastan**, J. Carbonell, and J. Klein-Seetharaman. A multi-task learning approach for virus-host interactions.

4. **O. Tastan**, J.J. Rose, S. Kollipara, S. Krishnamurthy, G.A. Gibson, T. Brosenitsch, R. Salter, S. Watkins, Z. Ambrose, C. Aiken, and J. Klein-Seetharaman. Mapping functionally important interactions between HIV-1 and human proteins.

Bibliography

[1] N. Legrand, A. Ploss, R. Balling, et al. Humanized mice for modeling human infectious disease: Challenges, progress, and outlook. *Cell Host Microbe*, 6(1):5–9, 2009.

[2] C. R. Parrish, E. C. Holmes, D. M. Morens, et al. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev*, 72(3):457–70, 2008.

[3] ECDC. Daily update, pandemic (h1n1) 2009. Technical report, European Center for Disease Prevention and Control, August 3, 2009 2009.

[4] News. Hiv vaccine failure prompts merck to halt trial. *Nature*, 449(7161):390, 2007.

[5] L. Menendez-Arias. Molecular basis of human immunodeficiency virus drug resistance: An update. *Antiviral Res*, 2009.

[6] J. Holland, K. Spindler, F. Horodyski, E. Grabau, S. Nichol, and S. VandePol. Rapid evolution of rna genomes. *Science*, 215(4540):1577–85, 1982.

[7] L. A. Shackelton, C. R. Parrish, U. Truyen, and E. C. Holmes. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc Natl Acad Sci U S A*, 102(2):379–84, 2005.

[8] E.D. Clercq. The design of drugs for hiv and hcv. *Nature Reviews Drug Discovery*, 6:1001–18, 2007.

[9] D. D. Ho and P. D. Bieniasz. Hiv-1 at 25. *Cell*, 133(4):561–5, 2008.

[10] R. J. Pomerantz. Reservoirs of human immunodeficiency virus type 1: The main obstacles to viral eradication. *Clin Infect Dis*, 34(1):91–7, 2002.

[11] S. Divito, T. L. Cherpes, and R. L. Hendricks. A triple entente: Virus, neurons, and cd8+ t cells maintain hsv-1 latency. *Immunol Res*, 36(1-3):119–26, 2006.

[12] B. Müller and H. Kräusslich. Antiviral strategies. In Hans-Georg Kräusslich and Ralf Bartenschlager, editors, *Antiviral Strategies*, volume 189 of *Handbook of Experimental Pharmacology*, pages 1–24. Springer Berlin Heidelberg, 2009.

[13] M.M. Lederman. Host-directed and immune-based therapies for human immunodeficiency virus infection. *Annals of Internal Med*, 122(3):218–222, 1995.

[14] S.L. Tan, G. Ganji, B. Paeper, S. Proll, and Katze MG. Systems biology and the host response to viral infection. *Nature Biotech*, 25(12):1383–9, 2007.

[15] A. Schwegmann and F. Brombacher. Host-directed drug targeting of factors hijacked by pathogens. *Sci Signal*, 1(29):re8, 2008.

[16] J. E. Garrus, U. K. von Schwedler, O. W. Pornillos, et al. Tsg101 and the vacuolar protein sorting pathway are essential for hiv-1 budding. *Cell*, 107(1):55–65, 2001.

[17] E. Wiertz, A. Hill, D. Tortorella, et al. Cytomegaloviruses use multiple mechanisms to elude the host immune response. *Immunol Lett*, 57(1-3):213–6, 1997.

[18] A. Alcami. Viral mimicry of cytokines, chemokines and their receptors. *Nat Rev Immunol*, 3(1):36–50, 2003.

[19] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, 3(4):e43, 2007.

[20] H. Jeong, S. P. Mason, A. L. Barabasi, et al. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 2001.

[21] Bruce Alberts. *Molecular biology of the cell*. Garland Science, New York, 5th edition, 2008.

[22] H. Pearson. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006.

[23] J. S. Mattick. The genetic signatures of noncoding rnas. *PLoS Genet*, 5(4):e1000459, 2009.

[24] J. S. Mattick and I. V. Makunin. Non-coding rna. *Hum Mol Genet*, 15 Spec No 1:R17–29, 2006.

[25] M. D. Adams, S. E. Celniker, R. A. Holt, et al. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–95, 2000.

[26] E. S. Lander, L. M. Linton, B. Birren, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[27] F. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–63, 1958.

[28] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.

[29] Harvey F. Lodish. *Molecular cell biology*. W.H. Freeman, New York, 6th edition, 2008.

[30] D. Beckett. Multilevel regulation of protein-protein interactions in biological circuitry. *Phys Biol*, 2(2):S67–73, 2005.

[31] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42, 2007.

[32] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–6, 1989.

[33] P. Obrdlik, M. El-Bakkoury, T. Hamacher, et al. K+ channel interactions detected by a genetic system optimized for systematic studies of membrane protein interactions. *Proc Natl Acad Sci U S A*, 101(33):12242–7, 2004.

[34] O. Puig, F. Caspary, G. Rigaut, et al. The tandem affinity purification (tap) method: A general procedure of protein complex purification. *Methods*, 24(3):218–29, 2001.

[35] B. Causier. Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrom Rev*, 23(5):350–67, 2004.

[36] R. B. Jones, A. Gordus, J. A. Krall, et al. A quantitative protein interaction network for the erbb receptors using protein microarrays. *Nature*, 439(7073):168–74, 2006.

[37] C. von Mering, R. Krause, B. Snel, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

[38] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–23, 2003.

[39] R. Mrowka, A. Patzak, and H. Herzel. Is there a bias in proteome research? *Genome Res*, 11(12):1971–3, 2001.

[40] M. Cornell, N. W. Paton, and S. G. Oliver. A critical and integrated view of the yeast interactome. *Comp Funct Genomics*, 5(5):382–402, 2004.

[41] M. A. Calderwood, K. Venkatesan, L. Xing, et al. Epstein-barr virus and virus human protein interaction maps. *Proc Natl Acad Sci U S A*, 104(18):7606–11, 2007.

[42] B. de Chassey, V. Navratil, L. Tafforeau, et al. Hepatitis c virus infection protein network. *Mol Syst Biol*, 4:230, 2008.

[43] P. L. Bartel and S. Fields. Analyzing protein-protein interactions using two-hybrid system. *Methods Enzymol*, 254:241–63, 1995.

[44] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–7, 2000.

[45] T. Ito, T. Chiba, R. Ozawa, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–74, 2001.

[46] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, et al. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–36, 2003.

[47] S. Li, C. M. Armstrong, N. Bertin, et al. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–3, 2004.

[48] J. F. Rual, K. Venkatesan, T. Hao, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8, 2005.

[49] P. Uetz, Y. A. Dong, C. Zeretzke, C. Atzler, et al. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239–42, 2006.

[50] H. Yu, P. Braun, M. A. Yildirim, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10, 2008.

[51] G. Butland, J. M. Peregrin-Alvarez, J. Li, et al. Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, 433(7025):531–7, 2005.

[52] A. C. Gingras, M. Gstaiger, B. Raught, et al. Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol*, 8(8):645–54, 2007.

[53] A. C. Gavin, P. Aloy, P. Grandi, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–6, 2006.

[54] N. J. Krogan, G. Cagney, H. Yu, et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–43, 2006.

[55] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–11, 1998.

[56] S. M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl. Duplexes of 21-nucleotide rnas mediate rna interference in cultured mammalian cells. *Nature*, 411(6836):494–8, 2001.

[57] M. Wassenegger. The role of the rnai machinery in heterochromatin formation. *Cell*, 122(1):13–6, 2005.

[58] D. Moazed. Small rnas in transcriptional gene silencing and genome defence. *Nature*, 457(7228):413–20, 2009.

[59] I. J. Macrae, K. Zhou, F. Li, et al. Structural basis for double-stranded rna processing by dicer. *Science*, 311(5758):195–8, 2006.

[60] C. Falschlehner, S. Steinbrink, G. Erdmann, and M. Boutros. High-throughput rnai screening to dissect cellular pathways: a how-to guide. *Biotechnol J*, 5(4):368–76, 2010.

[61] G. J. Hannon. Rna interference. *Nature*, 418(6894):244–51, 2002.

[62] S. P. Goff. Knockdown screens to knockout hiv-1. *Cell*, 135(3):417–20, 2008.

[63] M. M. Kulkarni, M. Booker, S. J. Silver, et al. Evidence of off-target effects associated with long dsrnas in drosophila melanogaster cell-based assays. *Nat Methods*, 3(10):833–8, 2006.

[64] N. J. Dimmock, Andrew Easton, and Keith Leppard. *Introduction to modern virology*. Blackwell Science, Oxford ; Malden, MA, 6th edition, 2007.

[65] N. Dimmock, A. Easten, and K. Leppard. *Introduction to Modern Virology*. John Wiley & Sons, Malden, MA, USA, 6th edition, 2006.

[66] J. Carter and Saunders V. *Virology: Principles and Applications*. CJohn Wiley & Sons, San Francisco, CA, USA, 1st edition, 2007.

[67] H. R. Gelderblom. Structure and classification of viruses. 1996. Gelderblom, Hans R. Book Chapter Galveston (TX).

[68] K.N. Ward, A. C. McCartney, and B. Thakker. *Notes on medical microbiology including virology, mycology and parasitology*. Churchill Livingstone, Edinburgh ; New York, 2nd edition, 2009.

[69] A. Trkola. Hiv-host interactions: Vital to the virus and key to its inhibition. *Curr Opin Microbiol*, 7(5):555–9, 2004.

[70] WHO. World health organization aids epidemic update 2007, joint united nations programme on hiv/aids (unaids) and who., 2007.

[71] S. Cherry. What have rnai screens taught us about viral-host interactions? *Curr Opin Microbiol*, 12(4):446–52, 2009.

[72] A. L. Brass, D. M. Dykxhoorn, Y. Benita, et al. Identification of host proteins required for hiv infection through a functional genomic screen. *Science*, 319(5865):921–6, 2008.

[73] H. Zhou, M. Xu, Q. Huang, et al. Genome-scale rnai screen for host factors required for hiv replication. *Cell Host Microbe*, 4(5):495–504, 2008.

[74] R. Kónig, Y. Zhou, D. Elleder, et al. Global analysis of host-pathogen interactions that regulate early-stage hiv-1 replication. *Cell*, 135(1):49–60, 2008.

[75] A. W. Tai, Y. Benita, L. F. Peng, et al. A functional genomic screen identifies cellular cofactors of hepatitis c virus replication. *Cell Host Microbe*, 5(3):298–307, 2009.

[76] T. I. Ng, H. Mo, T. Pilot-Matias, et al. Identification of host genes involved in hepatitis c virus replication by small interfering rna technology. *Hepatology*, 45(6):1413–21, 2007.

[77] M. N. Krishnan, A. Ng, B. Sukumaran, et al. Rna interference screen for human genes associated with west nile virus infection. *Nature*, 455(7210):242–5, 2008.

[78] L. Hao, A. Sakurai, T. Watanabe, et al. Drosophila rnai screen identifies host genes important for influenza virus replication. *Nature*, 454(7206):890–3, 2008.

[79] M. L. Yeung, L. Houzet, V. S. 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. *J Biol Chem*, 284(29):19463–73, 2009.

[80] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, et al. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, 1999.

[81] E. M. Marcotte, M. Pellegrini, H. L. Ng, et al. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–3, 1999.

[82] T. Dandekar, B. Snel, M. Huynen, et al. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–8, 1998.

[83] R. Overbeek, M. Fonstein, M. D'Souza, et al. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol*, 1(2):93–108, 1999.

[84] M. Y. Galperin and E. V. Koonin. Who's your neighbor? new computational approaches for functional genomics. *Nat Biotechnol*, 18(6):609–13, 2000.

[85] M. Huynen, B. Snel, 3rd Lathe, W., et al. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Res*, 10(8):1204–10, 2000.

[86] E. V. Koonin, Y. I. Wolf, and L. Aravind. Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome Res*, 11(2):240–52, 2001.

[87] S. A. Teichmann and M. M. Babu. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol*, 20(10):407–10; discussion 10, 2002.

[88] M. Pellegrini, E. M. Marcotte, M. J. Thompson, et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–8, 1999.

[89] D. Eisenberg, E. M. Marcotte, I. Xenarios, et al. Protein function in the post-genomic era. *Nature*, 405(6788):823–6, 2000.

[90] C. Ouzounis and N. Kyrpides. The emergence of major cellular processes in evolution. *FEBS Lett*, 390(2):119–23, 1996.

[91] P. Pagel, P. Wong, and D. Frishman. A domain interaction map based on phylogenetic profiling. *J Mol Biol*, 344(5):1331–46, 2004.

[92] A. Stein, A. Panjkovich, and P. Aloy. 3did update: Domain-domain and peptide-mediated interactions of known 3d structure. *Nucleic Acids Res*, 37(Database issue):D300–4, 2009.

[93] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–92, 2001.

[94] M. Deng, S. Mehta, F. Sun, et al. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540–8, 2002.

[95] W. K. Kim, J. Park, and J. K. Suh. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair. *Genome Inform*, 13:42–50, 2002.

[96] S. M. Gomez, W. S. Noble, and A. Rzhetsky. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, 19(15):1875–81, 2003.

[97] H. Yu, N. M. Luscombe, H. X. Lu, et al. Annotation transfer between genomes: Protein-protein interologs and protein-dna regulogs. *Genome Res*, 14(6):1107–18, 2004.

[98] R. Riley, C. Lee, C. Sabatti, and D. Eisenberg. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*, 6(10):R89, 2005.

[99] K. S. Guimaraes, R. Jothi, E. Zotenko, et al. Predicting domain-domain interactions using a parsimony approach. *Genome Biol*, 7(11):R104, 2006.

[100] M. Iqbal, A. A. Freitas, C. G. Johnson, and M. Vergassola. Message-passing algorithms for the prediction of protein domain interactions from protein-protein interaction data. *Bioinformatics*, 24(18):2064–70, 2008.

[101] D. Koller and F. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[102] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11(12):2120–6, 2001.

[103] B. Lehner and A. G. Fraser. A first-draft human protein-interaction map. *Genome Biol*, 5(9):R63, 2004.

[104] C. S. Goh, A. A. Bogan, M. Joachimiak, et al. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–93, 2000.

[105] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–14, 2001.

[106] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–901, 2002.

[107] P. Aloy and R. B. Russell. Interprets: Protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–2, 2003.

[108] L. Lu, H. Lu, and J. Skolnick. Multiprospector: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–64, 2002.

[109] J. Janin. Principles of protein-protein recognition from structure to thermodynamics. *Biochimie*, 77(7-8):497–505, 1995.

[110] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1):37–46, 2002.

[111] R. Jansen, H. Yu, D. Greenbaum, et al. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–53, 2003.

[112] R. Jansen and M. Gerstein. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol*, 7(5):535–45, 2004.

[113] A. Ben-Hur and W. S. Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1:S2, 2006.

[114] L. V. Zhang, S. L. Wong, O. D. King, and F. P. Roth. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, 5:38, 2004.

[115] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1:i38–46, 2005.

[116] D. R. Rhodes, S. A. Tomlins, S. Varambally, et al. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–9, 2005.

[117] N. Lin, B. Wu, R. Jansen, et al. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5:154, 2004.

[118] J. S. Bader, A. Chaudhuri, J. M. Rothberg, et al. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1):78–85, 2004.

[119] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500, 2006.

[120] H. W. Mewes, D. Frishman, U. Guldener, et al. Mips: A database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–4, 2002.

[121] T. S. Keshava Prasad, R. Goel, K. Kandasamy, et al. Human protein reference database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–72, 2009.

[122] T. P. Mohamed, J. G. Carbonell, and M. K. Ganapathiraju. Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, 11 Suppl 1:S57, 2010.

[123] M. D. Dyer, T. M. Murali, and B. W. Sobral. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics*, 23(13):i159–66, 2007.

[124] P. Evans, W. Dampier, L. Ungar, et al. Prediction of hiv-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med Genomics*, 2:27, 2009.

[125] O. Tastan, Y. Qi, J. G. Carbonell, et al. Prediction of interactions between hiv-1 and human proteins by information integration. *Pac Symp Biocomput*, pages 516–27, 2009.

[126] F. P. Davis, D. T. Barkan, N. Eswar, et al. Host pathogen protein interactions predicted by comparative modeling. *Protein Sci*, 16(12):2585–96, 2007.

[127] F. P. Davis, H. Braberg, M. Y. Shen, et al. Protein complex compositions predicted by structural similarity. *Nucleic Acids Res*, 34(10):2943–52, 2006.

[128] S. A. Lee, C. H. Chan, C. H. Tsai, et al. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*, 9 Suppl 12:S11, 2008.

[129] N. Tyagi, O. Krishnadev, and N. Srinivasan. Prediction of protein-protein interactions between helicobacter pylori and a human host. *Mol Biosyst*, 5(12):1630–5, 2009.

[130] O. Krishnadev and N. Srinivasan. Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int J Biol Macromol*, 48(4):613–9, 2011.

[131] J. M. Doolittle and S. M. Gomez. Structural similarity-based predictions of protein interactions between hiv-1 and homo sapiens. *Virol J*, 7:82, 2010.

[132] H. Huang and J. S. Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3):372–8, 2009.

[133] O. Tastan, S. Jäeger, J. Carbonell, N. Krogan, and J. Klein-Seetharaman. Prediction of hiv-1 human interactions - an improved framework. *In submission*, 2010.

[134] R. G. Ptak, W. Fu, B. E. Sanders-Beer, et al. Cataloguing the hiv type 1 human protein interaction network. *AIDS Res Hum Retroviruses*, 24(12):1497–502, 2008.

[135] W. Fu, B. E. Sanders-Beer, K. S. Katz, et al. Human immunodeficiency virus type 1, human protein interaction database at ncbi. *Nucleic Acids Res*, 37(Database issue):D417–22, 2009.

[136] J. P. Mackay, M. Sunde, J. A. Lowry, et al. Protein interactions: Is seeing believing? *Trends Biochem Sci*, 32(12):530–1, 2007.

[137] Joel P. Mackay, Margaret Sunde, Jason A. Lowry, et al. Response to chatr-aryamontri et al.: Protein interactions: To believe or not to believe? *Trends in Biochemical Sciences*, 33(6):242–3, 2008.

[138] A. Chatr-Aryamontri, A. Ceol, L. Licata, et al. Protein interactions: Integration leads to belief. *Trends Biochem Sci*, 33(6):241–2; author reply 2–3, 2008.

[139] G. R. Welch. The 'fuzzy' interactome. *Trends Biochem Sci*, 34(1):1–2; author reply 3, 2009.

[140] M. Seringhaus and M. Gerstein. Manually structured digital abstracts: A scaffold for automatic text mining. *FEBS Lett*, 582(8):1170, 2008.

[141] A. Ceol, A. Chatr-Aryamontri, L. Licata, et al. Linking entries in protein interaction database to structured text: The febs letters experiment. *FEBS Lett*, 582(8):1171–7, 2008.

[142] M. E. Cusick, H. Yu, A. Smolyar, et al. Literature-curated protein interaction datasets. *Nat Methods*, 6(1):39–46, 2009.

[143] O. Tastan, J. Carbonell, and J. Klein-Seetharaman. A multi-task learning approach for virus-host interactions. *In submission*, 2010.

[144] S. P. Goff. Host factors exploited by retroviruses. *Nat Rev Microbiol*, 5(4):253–63, 2007.

[145] L Breiman. Random forests. *Mach. Learn.*, 45:5–32, October 2001.

[146] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997,.

[147] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[148] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, pages 81–106, 1986.

[149] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. CRC Press, 2010.

[150] B. Tso and P.M. Mather. *Classification methods for remotely sensed data 2nd edition*. Taylor & Francis Group LLC, Florida, 2nd edition, 2001.

[151] P. Shannon, A. Markiel, O. Ozier, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

[152] M. Ashburner, C. A. Ball, J. A. Blake, et al. Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000.

[153] C. Brooksbank, G. Cameron, and J. Thornton. The european bioinformatics institute's data resources: Towards systems biology. *Nucleic Acids Res*, 33(Database issue):D46–53, 2005.

[154] P. Puntervoll, R. Linding, C. Gemund, et al. Elm server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*, 31(13):3625–30, 2003.

[155] R. Apweiler, T. K. Attwood, A. Bairoch, et al. The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 29(1):37–40, 2001.

[156] T. Barrett, D. B. Troup, S. E. Wilhite, et al. Ncbi geo: Mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35(Database issue):D760–5, 2007.

[157] S. Peri, J. D. Navarro, T. Z. Kristiansen, et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32(Database issue):D497–501, 2004.

[158] C. H. Wu, R. Apweiler, A. Bairoch, et al. The universal protein resource (uniprot): An expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–91, 2006.

[159] Los alomos hiv sequence database, 2008.

[160] T. Y. Lee, H. D. Huang, J. H. Hung, et al. Dbptm: An information repository of protein post-translational modification. *Nucleic Acids Res*, 34(Database issue):D622–7, 2006.

[161] Z. Du, L. Li, C. F. Chen, et al. G-sesame: Web tools for go-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res*, 37(Web Server issue):W345–9, 2009.

[162] L C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(40):35–41, 1977.

[163] M. D. Hyrcza, C. Kovacs, M. Loutfy, et al. Distinct transcriptional profiles in ex vivo cd4+ and cd8+ t cells are established early in human immunodeficiency virus type 1 infection and are characterized by a chronic interferon response as well as extensive transcriptional changes in cd8+ t cells. *J Virol*, 81(7):3477–86, 2007.

[164] E. Masliah, E. S. Roberts, D. Langford, et al. Patterns of gene dysregulation in the frontal cortex of patients with hiv encephalitis. *J Neuroimmunol*, 157(1-2):163–75, 2004.

[165] M. T. Vahey, M. E. Nau, L. L. Jagodzinski, et al. Impact of viral infection on the gene expression profiles of proliferating normal human peripheral blood mononuclear cells infected with hiv type 1 rf. *AIDS Res Hum Retroviruses*, 18(3):179–92, 2002.

[166] J.A. Levy. *HIV and the pathogenesis of AIDS*. ASM Press, Washington, DC, USA, 3rd edition, 2007.

[167] S. F. Altschul, W. Gish, W. Miller, et al. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[168] M. Matsubara, T. Jing, K. Kawamura, et al. Myristoyl moiety of hiv nef is involved in regulation of the interaction with calmodulin in vivo. *Protein Sci*, 14(2):494–503, 2005.

[169] UniProt Knowledgebase. Controlled vocabulary of posttranslational modifications. `http://www.uniprot.org/docs/ptmlist`, February 2008.

[170] R. Baeza-Yates and Berthier de Araújo Neto Ribeiro. *Modern information retrieval*. ACM Press ; Addison-Wesley, New York Harlow, England, 1999.

[171] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson. Ontologizer 2.0–a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1, 2008.

[172] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, et al. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–2, 2002.

[173] H. Liang and W. H. Li. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet*, 23(8):375–8, 2007.

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

[175] D. E. Ott. Cellular proteins detected in hiv-1. *Rev Med Virol*, 18(3):159–75, 2008.

[176] K. Zander, M. P. Sherman, U. Tessmer, K. Bruns, and others. Cyclophilin a interacts with hiv-1 vpr and is required for its functional expression. *J Biol Chem*, 278(44):43202–13, 2003.

[177] J. Luban, K. L. Bossolt, E. K. Franke, G. V. Kalpana, and S. P. Goff. Human immunodeficiency virus type 1 gag protein binds to cyclophilins a and b. *Cell*, 73(6):1067–78, 1993.

[178] B. J. Kovaleski, R. Kennedy, M. K. Hong, S. A. Datta, L. Kleiman, A. Rein, and K. Musier-Forsyth. In vitro characterization of the interaction between hiv-1 gag and human lysyl-trna synthetase. *J Biol Chem*, 281(28):19449–56, 2006.

[179] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H. W. Mewes. Corum: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res*, 38(Database issue):D497–501, 2009.

[180] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39(Database issue):D685–90, 2010.

[181] A. Chatr-aryamontri, A. Ceol, D. Peluso, A. Nardozza, S. Panni, F. Sacco, M. Tinti, A. Smolyar, L. Castagnoli, M. Vidal, M. E. Cusick, and G. Cesareni. Virusmint: a viral protein interaction database. *Nucleic Acids Res*, 37(Database issue):D669–73, 2009.

[182] I. Singh, O. Tastan, and J. Klein-Seetharaman. Comparison of virus interactions with human signal transduction pathways. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, BCB '10, pages 17–24, 2010.

[183] A. Zamborlini, A. Coiffic, G. Beauclair, et al. Impairment of human immunodeficiency virus type-1 integrase sumoylation correlates with an early replication defect. *J Biol Chem*, 2011.

[184] R. Geiss-Friedlander and F. Melchior. Concepts in sumoylation: a decade on. *Nat Rev Mol Cell Biol*, 8(12):947–56, 2007.

[185] R. Boggio and S. Chiocca. Viruses and sumoylation: recent highlights. *Curr Opin Microbiol*, 9(4):430–6, 2006.

[186] C. Gurer, L. Berthoux, and J. Luban. Covalent modification of human immunodeficiency virus type 1 p6 by sumo-1. *J Virol*, 79(2):910–7, 2005.

[187] B. J. Breitkreutz, C. Stark, T. Reguly, et al. The biogrid interaction database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D637–40, 2008.

[188] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, et al. Mint: The molecular interaction database. *Nucleic Acids Res*, 35(Database issue):D572–4, 2007.

[189] I. Xenarios, E. Fernandez, L. Salwinski, et al. Dip: The database of interacting proteins: 2001 update. *Nucleic Acids Res*, 29(1):239–41, 2001.

[190] S. Kerrien, Y. Alam-Faruque, B. Aranda, et al. Intact–open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue):D561–5, 2007.

[191] G. D. Bader, I. Donaldson, C. Wolting, et al. Bind–the biomolecular interaction network database. *Nucleic Acids Res*, 29(1):242–5, 2001.

[192] Editorial. Maturing interactions. *Nature Methods*, 6(1):2, 2009.

[193] Mint database scoring function, 2009.

[194] L. von Ahn, B. Maurer, C. McMillen, et al. Recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–8, 2008.

[195] D. Whitford. Hired guns on the cheap. *Fortune Small Business*, (March 1, 2007), 2007.

[196] S. Sheng Victor, Provost Foster, and G. Ipeirotis Panagiotis. Get another label? improving data quality and data mining using multiple, noisy labelers, 2008.

[197] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1Ű38, 1977.

[198] C. B. Do and S. Batzoglou. What is the expectation maximization algorithm? *Nat Biotechnol*, 26(8):897–9, 2008.

[199] S. Urata, H. Yokosawa, and J. Yasuda. Regulation of htlv-1 gag budding by vps4a, vps4b, and aip1/alix. *Virol J*, 4:66, 2007.

[200] S. Balakrishnan, O. Tastan, J. Carbonell, and J. Klein-Seetharaman. Alternative paths in hiv-1 targeted human signal transduction pathways. *BMC Genomics*, 10 Suppl 3:S30, 2009.

[201] D. T. O'Hagan and N. M. Valiante. Recent advances in the discovery and delivery of vaccine adjuvants. *Nat Rev Drug Discov*, 2(9):727–35, 2003.

[202] Yi-Wei C. and Chih-Jen L. Combining svms with various feature selection strategies. In Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti Zadeh, editors, *Feature extraction, foundations and applications*, Studies in fuzziness and soft computing. Springer, 2006.