ANDRÉS BECERRA SANDOVAL

CHARACTERIZING VIRUSES MIMICRY MECHANISMS

SUPERVISORS

PEDRO A. MORENO TOVAR & VICTOR A. BUCHELLI GUERRERO

DOCTORADO EN INGENIERÍA
ÉNFASIS EN CIENCIAS DE LA COMPUTACIÓN
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
FACULTAD DE INGENIERÍA

UNIVERSIDAD DEL VALLE

CHARACTERIZING VIRUSES MIMICRY MECHANISMS

WITH PROTEIN-PROTEIN INTERACTIONS AND SHORT LINEAR MOTIFS

ANDRÉS BECERRA SANDOVAL

ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
FACULTAD DE INGENIERÍA

September - 2016 – final version

SUPERVISORS
Pedro A. Moreno Tovar
Victor A. Buchelli Guerrero

EVALUATION COMITTEE
Juan Manuel Anzola
Aydin Tozeren
Felipe García Vallejo

PUBLIC DEFENSE:
Santiago de Cali, august 26 - 2016

To Sandra:
For her inconditional support

To my brother, mother and father Fernando, Elena y Argemiro:
for their support

To mi sons Nicolás y Alejandro:
for their patience

ABSTRACT

Viruses are obligate intracellular parasites predominant in all domains of life. Viral infections cause diseases and death to humans and edible organisms like plants and cattle. Viral short genomes encode subversion mechanisms of the host cell subsystems. These subversion mechanisms are based on virus-host protein-protein interactions.

The descent of next generation sequencing costs has spurred an exponential growth of the number of viral genomes in bioinformatic databases. In contrast, the cost of experimental high-throughput techniques for virus-host protein-protein determination is not decreasing at the same rate. Although the number of virus-host protein-protein interactions has been growing in databases too, it is still low in order to analyze viral subversion mechanisms with system biology approaches.

There is a need of computational virus-host protein-protein interaction prediction methods. However, the number of viral and host protein 3D structures is too small in order to predict virus-host protein-protein interactions with structural methods.

The prediction of virus-host protein-protein interactions has been carried out mostly with machine learning classifiers. However, the classifiers do not reveal in high-level why the interactions were inferred and are sensitive to the training set quality. For virus-host protein-protein interactions there is still not a gold standard data set for validation of interactions.

This thesis is about predicting virus-host interactions mediated by protein short linear motifs. These interactions are more predominant than virus-host domain-domain interactions and are used by several viruses. Additionally, the inference of this kind of interactions is guided by biological hypothesis like the conservation of motifs and localization of motifs in protein disordered regions.

The result of this thesis is a computational method and platform for predicting interactions between host and viral proteins mediated by linear motifs. This candidate interactions are used to study common viral attack strategies in a particular human subsystem, the protein-synthesis machinery. Nevertheless, the methods developed can be used with any subsystem like interferon and apoptosis.

RESUMEN

Los virus son parásitos intracelulares obligados que predominan en todos los dominios de la vida. Las infecciones virales causan enfermedades y muertes a seres humanos y a organismos que sirven de alimento como plantas y ganado. Los cortos genomas virales codifican mecanismos de subversión de las células hospederas. Estos mecanismos están basados en interacciones proteína-proteína entre virus y hospedero.

La disminución de los costos de la secuenciación de nueva generación ha impulsado un crecimiento exponencial del número de genomas virales en las bases de datos bioinformáticas. En contraste, el costo de las técnicas experimentales para la determinación de interacciones proteína-proteína entre virus y hospedero no está descendiendo a la misma velocidad. Aunque el número de interacciones proteína-proteína entre virus y hospedero ha crecido en las bases de datos, todavía es bajo para analizar los mecanismos de subversión viral con enfoques de biología de sistemas.

Hay una necesidad de métodos computacionales de predicción de interacciones proteína-proteína entre virus y hospedero. Sin embargo, el número de estructuras 3D de proteínas virales y de hospederos es muy pequeño para realizar la predicción con métodos estructurales.

La predicción de interacciones proteína-proteína entre virus y hospedero ha sido realizada principalmente con clasificadores basados en aprendizaje automático. No obstante, los clasificadores desarrollados no revelan en alto nivel por qué se infirieron las interacciones y son sensitivos a la calidad del conjunto de datos de aprendizaje. Para interacciones proteína-proteína entre virus y hospedero todavía no hay un conjunto de datos de validación de alta calidad.

Esta tesis trata sobre la predicción de interacciones proteína-proteína entre virus y hospedero mediadas por motivos lineales cortos. Estas interacciones predominan más que las interacciones dominio-dominio entre virus y hospedero; además son usadas por varios virus. La inferencia de este tipo de interacciones se sustenta en hipótesis biológicas como la conservación de los motivos y su localización en regiones proteínicas desordenadas.

El resultado de ésta tesis es un método y una plataforma computacional para predecir interacciones proteína-proteína entre virus y hospederos mediadas por motivos lineales. Las interacciones candidatas obtenidas son usadas para estudiar un subsistema particular humano, las proteínas que hacen síntesis de proteínas. Sin embargo, los métodos desarrollados pueden ser usados con cualquier otro subsistema como el interferon y el de apoptosis.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Computational analysis of the linear motif mediated subversion of the human protein synthesis machinery, BIOTECHNO 2016: The Eighth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies.

ACKNOWLEDGMENTS

To professors Irene Tischer, Angel García Baños, Victor Buchelli, Pedro Moreno in the Escuela de ingeniería de sistemas y computación (EISC) in the Universidad del Valle.

To professor Aydin Tozeren for receiving me in his lab at the Biomedical Engineering Department at the Drexel university.

To the evaluation comitee for their time and valuable comments to improve this thesis, Doctors Juan Manuel Anzola, Aydin Tozeren and Felipe García Vallejo.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

CATH Classification of protein structures downloaded from the Protein Data Bank (Class, Architecture, Topology,Homologous superfamily)

CSV Comma Separated Value format

DDI Domain-Domain Interaction

DMI Domain-Motif Interaction

DNA Deoxyribonucleic Acid

EBV Epstein-Bar Virus

ELM Eukaryotic Linear Motifs

GO Gene Ontology

GWAS Genome Wide Association Study(ies)

HIV-1 Human Immunodeficiency Virus 1

HPSM Host Protein Synthesis Machinery

ICTV International Committee on Taxonomy of Viruses

KEGG Kyoto Encyclopedia of Genes and Genomes

LUCA Last Universal Common Ancestor

mRNA messenger RNA

NIAID National Institute of Allergy and Infectious Diseases (USA)

PDB Protein Data Bank

PHI Pathogen-Host (protein-protein) interactions

PIN(s) Protein Interaction Network(s)

PPIs Protein-Protein Interactions

PTM Post-Translational Modification

RNA Ribonucleic Acid

RPGdb Ribosomal Protein Gene database

SCOP Structural Classification of Proteins

SLiMs Short Linear Motifs

SNP Single Nucleotide Polymorphisms

YTH Yeast-Two-Hibrid assay

VHPPIs Virus-Host Protein-Protein Interactions

1 INTRODUCTION

BACKGROUND

Protein-protein interactions are used to understand the biological organization of an organism by creating models like metabolic, regulation, signaling and interaction networks. In the same way, virus-host protein-protein interactions (VHPPIs) are used to create models aimed at understand viral pathogenesis.

The reason VHPPIs are inputs to formulate pathogenesis models is that viruses interact with their hosts in order to disrupt or modulate pathways achieving goals like evading the complement system [1], modulating the cytokine system [3] and abrogating apoptosis [89].

Some of these protein-protein interactions (PPIs) are based on viral mimicry. A viral protein (VP) resembling a host protein (HP) might interact with other host proteins, the ones that interact with HP.

Classic definition of molecular mimicry

Oldstone defined molecular mimicry in reference [151] as:

> The sharing of a linear amino acid sequence or a conformation fit between a microbe and a host 'self' determinant is the initial stage of molecular mimicry. Autoimmunity may occur if the host immune response against the microbe cross-reacts with the host's 'self' sequence and if the host sequence comprises a biologically important domain.

There is a balance between the pathogen virulence and the host immune system response. This response consist in the generation of antibodies to bind pathogen proteins. The antibody-bound pathogen proteins are degraded by the immuno-proteasome into peptides. These peptides are presented by proteins in the Major histocompatibility complex (MHC) to appropriate immune T-cells. The T-cells decide if cells are infected by pathogens and initiate apoptosis. MHC class I proteins present cytosolic peptides and MHC class II proteins display extracellular peptides. The immune system can go out of control when the peptides are identical to a host protein component and start to attack host cells unfolding an autoimmune disease like insulin-dependent diabetes and Guillain-Barré syndrome.

An individual immune system response depends on its genetic predisposition and environmental factors like pathogen infection history. The results can be: 1) initiation of autoimmune processes when

pathogens have an epitope (a part of a pathogenic protein that binds to an antibody) identical to a host protein component; 2) acceleration of ongoing autoimmune processes by subsequent infections, when the different pathogens have very similar epitopes, and these epitopes have similarity to host protein components; and 3) abrogation of autoimmunity by subsequent infections, through different mechanisms like inflammation that cause hyperactivation of lymphocytes, that in turn, may diminish the number of aggressive T cells [33].

Molecular mimicry definition used in this thesis

The notion of mimicry we are going to use throughout the thesis is the one presented in reference [124]:

> We refer here to molecular mimicry as the display of any structure by the parasite that (i) resembles structures of the host at the molecular level and (ii) confers a benefit to the parasite because of this resemblance.

Or the similar one presented in reference [66]:

> We define mimics as pathogen-encoded factors that resemble host factors in order to co-opt or disrupt host functions to the pathogen's advantage.

From now on, viral mimicry is considered as any molecular resemblance to host proteins based on sequence or structural similarity that confers the virus ways to hijack or disrupt host pathways. This definition of molecular mimicry is adopted because it serves the purpose of finding virus-host interactions more than the classical one. The classical is more suited to investigate autoimmune disorders though.

This thesis is centered in the prediction VHPPIs based on mimicked protein portions known as short linear motifs (SLiMs) and the analysis of the interactions obtained to find candidate viral strategies based in this kind of mimicry.

JUSTIFICATION

Systems biology approaches use VHPPIs to formulate hypothesis about viral infection mechanisms [59]. The hypothesis obtained are aimed at developing antivirals and vaccines [45, 214, 53, 125].

However, the scarcity of VHPPIs with experimental evidence support is an obstacle to system biology approaches [60]. This lack of data has fostered the development of VHPPI prediction methods.

Prediction methods use sequences as inputs because the low cost of next generation sequencing technologies has allowed to sequence a big number of host and pathogen genomes. Furthermore, the genome sequencing costs are declining as new technologies emerge.

On the other hand, there are few 3D protein structures determined for hosts and viruses in the protein data bank (PDB), the main repository of protein structural information [169]. These 3D structures could be used as input for structure-based PPI prediction methods like the one implemented in reference [54].

The abundance of sequenced genomes and viral protein sequences have allowed the development of several bioinformatic methods to predict VHPPIs [11, 39, 54, 63, 65, 149, 164]. Most of the prediction methods developed are machine learning classifiers like random forests [208] and support vector machines [11, 39, 65].

Most of these classifiers use protein sequences and other features like gene ontology (GO) function, gene co-expression, homology with similar organisms and cellular localization among others.

All the classifiers face two problems: the missing data problem, that all the proteins have no data for some features [110]; and the negative training data problem, the adequate choosing of negative examples — pairs of non-interacting proteins– for training [13]. Indeed, the good performance of the classifiers reported in the literature can be a consequence of the biased selection of negative examples [13].

Another problem with machine learning methods lies in the difficulty to interpret biologically the inference of the results of a properly trained classifier. The inference of VHPPIs ends up encoded in neuron weights or support vector values, there is no a high level explanation of the inference process.

The lack of 3D structures for structural prediction methods and protein features for machine learning classifiers introduced above underscore the importance of methods based only on protein sequences which is the case for the method developed in this thesis.

Another property of the method developed is that it leads to a straightforward biological explanation of the interactions inference process. Indeed, the inference of interactions based on SLiMs uses criteria as SLiM conservation, SLiM localization in protein structurally disordered regions and difficulty to form SLiMs by pure chance that have a biological foundation in evolution, protein structure and probability, respectively.

Recently, the role of SLiMs has been studied in a wide set of viruses [87]. The authors conclude that viruses use extensively SLiMs as means to interact with host proteins. Also, human proteins targeted by viruses have a high number of SLiMs [88].

An analysis of VHPPIs with experimental support shows that if the interactions are classified in domain-motif interactions (DMI) and domain-domain interactions (DDI), DMI are the predominant ones. Furthermore, DMI are used by several viruses while DDI are virus-specific [88].

For the reasons introduced above, the viral use of SLiMs and DMI predominance/reuse, we argue that virus-host interactions mediated by SLiMs can be used to analyze common viral attack strategies.

As there is no a gold-standard data set for VHPPIs to validate a prediction method we use the human immunodeficiency virus 1 (HIV-1) to analyze the VHPPIs deduced used SLiMs.

As HIV-1 is a human retrovirus that disables the immune system, mutates extraordinarily fast [38] and infects so many people around the world it has been extensively investigated. This makes it the virus with more bioinformatic data available, with the USA National Institute of Allergy and Infectious Diseases (NIAID) databases for sequences and alignments [111] and for interactions with human proteins [79].

Viruses must subvert different subsystems to infect the host cell and replicate. Although viruses have an extraordinary variability, all must subvert the host protein synthesis machinery (HPSM) in order to translate viral proteins. That is the reason we choose this particular subsystem for study of common viral infection strategies. We select a group of human viruses with available sequences targeting the host protein-synthesis machinery as a case study.

PREVIOUS WORK

Eukaryotic organisms have an estimated huge number of SLiM instances in their proteins, more than a million in humans[191]. These SLiMs are used as dynamic interaction anchors than can appear and disappear with small mutations and tune host protein-protein interaction networks (PPIN) [143].

The first investigation to predict VHPPI with base on SLiMs was conducted by Evans et al. with HIV-1 and human proteins [68]. The authors use SLiM conservation as a criterion to filter SLiM instances before inferring the interactions.

A massive study of SLiM use with 2208 different viral genomes was conducted by Hagai et al. [87]. The authors find that SLiM instances are prevalent in viral protein sequences, even more in eukaryotic than prokaryotic organisms. They propose localization in protein disordered regions and difficulty to form by pure chance as criteria to filter SLiM instances.

Furthermore, they find that SLiM instances emerge in viruses by convergent evolution predominantly and not by horizontal gene transfer [87]. This suggests that unrelated viruses can use the same SLiMs in host proteins to bridge new interactions or disrupt existing pathways, supporting our aim of finding common viral infection mechanisms based on mimicry.

SCOPE

Although protein-protein interactions take place on the basis of the interacting proteins 3D structure, we only use sequence information in this thesis to predict interactions. We do not use 3D protein information because the number of viral protein structures in PDB is small. This decision prevent us from including the role of water in the interactions that can be very influential facilitating the interactions and conditioning the folding of interacting proteins into secondary and tertiary structures.

STRUCTURE OF THE DOCUMENT

The organization of this thesis is as follows. In Chapter 2 we present the objectives, in part i we introduce some concepts in virus molecular biology and for HIV-1 in particular, Chapter 3, and the state of the art on bioinformatics and systems biology for studying viral infections, Chapter 4.

In Part ii we present the methods used for developing the method with HIV-1 data, Chapter 5 , and for analyzing the viral subversion of the HPSM, Chapter 6.

In Part iii we present the results obtained for human-HIV-1 interactions, Chapter 7, and for the analysis of the HPSM, chatper 8.

Finally, in Part iv we present the conclusions of the thesis and directions for further research.

2

OBJECTIVES

2.1 GENERAL OBJECTIVE

The main objective of this thesis is to develop a computational model and platform for predicting virus-host protein-protein interactions mediated by short linear motifs and analyze viruses subversion strategies of the host protein-synthesis machinery based on short linear motif mimicry.

2.2 SPECIFIC OBJECTIVES

1. Select proper data sets to develop and test the model.

   We choose HIV-1 sequences and interactions with human proteins to calibrate the method. For the host protein-synthesis machinery, the subset of human proteins was constructed following reference [201] and the Uniprot database [12].

2. Generalize the approach of Evans et al. to any virus with sequenced genome in order to predict interactions based on common SLiMs between human and viral proteins.

   We generalize the Evans et al. approach for any eukaryotic virus using the ELM database SLiM-domain associations [51], and Pfam domain-protein associations [72] to get the SLiM-protein associations [68]. We implement algorithms to count and measure the conservation of SLiMs in the viral sequences.

3. Extend the approach of Evans et al. with the criteria of protein structural disorder and randomization of proteins to find SLiMs hard to form by pure chance.

   We extend the criterion of conserved common SLiMs proposed by Evans et al with the notion of SLiM localization in protein disordered regions using a protein disorder prediction strategy based on protein sequence only. We also implement the notion of difficulty of finding motifs by pure chance in protein randomized sets proposed by Hagai et al. [87].

4. Conduct a case study of the computational model proposed on a specific virus.

   We calibrate the method developed by comparing how the three criteria of SLiM filtering – conservation, localization in disordered regions, difficulty to form by pure chance – perform at

predicting virus-host protein-protein interactions. We use HIV-1 to analyze the results because is the virus with more abundant information.

5. Conduct a case study of the analysis with several human viruses on the HPSM.

   We focus on the HPSM subsystem because all viruses subvert it to translate their proteins. We focus on human as host to obtain insights into common viral infection strategies. These insights could be relevant to prevent viral infectious diseases.

6. Create a computational tool, implementing the developed model and the corresponding algorithms, to be used by bioinformaticians to study viral infections.

   We develop a set of scripts written in the Python programming language, using the libraries Biopython, Pandas and NetworkX. The scripts are general enough to be used with any eukaryotic virus, using the Fasta format for protein sequences and Uniprot identifiers for proteins.

   The resulting interactions are generated in CSV (comma separated value) spreadsheets and the resulting virus host protein-protein interaction networks in the Graphml format that can be analyzed with tools like Cytoscape [177].

Part I

STATE OF THE ART

3

VIRUS MOLECULAR BIOLOGY WITH EMPHASIS ON HIV-1

3.1 BACKGROUND

Organisms on earth can be classified on two types: capsid-encoding organisms (viruses) and ribosome-encoding organisms in the three domains of life, archaea, bacteria and eukarya [108]. Viruses are the smallest known forms of life: obligate intracellular parasites lacking metabolic processes and protein-synthesis machinery that must replicate by infecting and subverting ribosome-encoding cells. Viral genomes can be Deoxyribonucleic Acid (DNA) or Ribonucleic Acid (RNA).

Although viruses origin is unknown at the moment, the presence of viruses in all the ribosome-encoding organisms and the remarkable differences between the viral lineages might suggest that viruses are polyphyletic. Although there is statistical evidence that ribosome-encoding organisms have a common descent from "LUCA" (the hypothetical Last Universal Common Ancestor) [190], different types of viruses appear to have evolved several times and co-evolved with the first cells [109].

Indeed, there are three hypothesis for the origin of viruses: 1) the progressive, 2) the regresive and 3) the virus-first. In the progressive hypothesis (1) genetic mobile elements like plasmids, transposons and retrotransposons are the ancestors of viruses and acquired proteins that gave them the capability of moving between cells; the similarity between retrotransposons and retroviruses supports this hypothesis as the evidence for acquisition of cellular genes by Nucleocytoplasmic large DNA viruses [71]. In the regresive hypothesis (2) viral ancestors were cellular parasitic organisms that lost parts of their genomes through time. In the virus-first hypothesis (3) viruses formed concurrently with other selfish agents and the first cells originating the three extant viral lineages (archaea, bacterya, eukarya) [109].

Viruses have two stages: as a virion (particle) and infecting a host cell. Virions are formed by self-assembly of components in the infected host cell. The infected host cell is commandeered by the virus genome products to synthesize the components that can be assembled in new virions. The changes that a virion undergoes to command a host cell are called the infectious cycle.

The remarkable tasks a virus accomplishes in the infectious cycle are carried out with the smaller genomes known on earth. Consider

that HIV-1, one of the most dangerous and studied viruses, encodes around 20 proteins and 30 micro RNAs (miRNA) [118, 188]. Although virus lineages are remarkably different in their replication schemes, there are structural similarities in their particles (virions) and in the way the infection processes take place.

Although the molecular biology of viruses is complex and extensive this chapter summarize the main points that will be needed in the following chapters with an emphasis in HIV-1 as a concrete example.

3.2 DETECTION AND CLASSIFICATION

3.2.1 *Detection*

Koch postulates, as presented in reference [78] were used as a way to confirm that a a pathogen causes a disease:

> (i) The parasite occurs in every case of the disease in question and under circumstances which can account for the pathological changes and clinical course of the disease.
>
> (ii) The parasite occurs in no other disease as a fortuitous and non-pathogenic parasite.
>
> (iii) After being fully isolated from the body and repeatedly grown in pure culture, the parasite can induce the disease anew.

Advances in molecular biology technology have changed this definition to the one given by Falkow [70], the one based on next generation sequencing technologies given by Fredericks et al. [78], and the one based on meta-genomics given by Mokili et al. [136].

In clinical practice viruses like HIV-1 are detected with less expensive tests for *known* antibodies produced by the host in response to the virus. Commonly used tests are the Western blot and indirect immunofluorescence assays.

3.2.2 *Classification*

There are several classification systems based on different criteria. One of the most used is the Baltimore classification system, based on the method viruses use to synthesize mRNA for replication: using DNA or RNA as genetic content, in single (**ss**) or double strand (**ds**) conformation, using a positive (**+**) or negative sense (**-**) for reading the nucleotides, and, if the viruses use reverse transcription (**RT**) as replication mechanism. The seven classes are presented below:

Table 1: Baltimore Classification

Type	mRNA synthesis system
I	dsDNA
II	ssDNA
III	dsRNA
IV	(+)ssRNA
V	(-)ssRNA
VI	ssRNA-RT
VII	dsRNA-RT

The International Committee on Taxonomy of Viruses (ICTV) uses a hierarchy of order, family, sub-family, genus and species to classify viruses. In the 2014 revision there are 7 orders and 78 families not assigned to any order.

HIV-1, the human immunodeficiency virus 1, is classified in the retroviridae (retrovirus) family, orthoretrovirinae subfamily, lentivirus (long incubation period) genus. In the Baltimore classification HIV-1 is in group VI (ssRNA-RT).

3.3 GENOMES

Viral genomes are very complex. Besides the possibilities of being DNA or RNA, they can be DNA with short segments of RNA, DNA or RNA with covalently attached protein. The orientation can be single stranded in the positive (+) and negative (–) sense or even ambisense. The overall shape can be linear, circular, segmented or gapped [74].

Viral genomes use several complex encoding strategies like the production of multiple sub-genomic mRNAs, mRNA splicing, RNA editing, and nested transcription units. All these methods compress genomic information. Additionally, post-transcriptional mechanisms like polyprotein synthesis, leaky scanning, suppression of termination, and ribosomal frame-shifting are used in viruses to expand the information encoded in their short genomes.

The 3D structure of single stranded RNA viral genomes contain substructures that regulate protein synthesis and the infectious cycle. Some substructures known are internal ribosome entry sites (IRES), packaging signals, pseudo-knots, tRNA mimics, ribosomal frame-shift motifs, and cis-regulatory elements. In HIV-1 these structures have functions like the activation of transcription, initiation of reverse transcription, facilitation of genomic dimerization, direction of virion packaging, manipulation of reading frames, regulation of RNA nuclear export, signaling polyadenylation, and interaction with viral and host proteins [205].

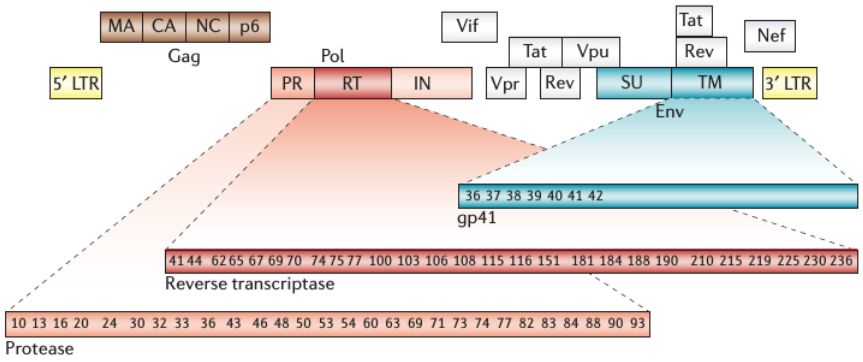


Figure 1: HIV-1 genome structure.
Image from Lengauer, T. and Sing, T., "Bioinformatics-assisted anti-HIV therapy", Nat. Rev. Microbiol. (2006), 790--797. Copyright Nature reviews. Microbiology ©2006

3.3.1 *HIV-1 genome*

HIV-1 genome has three reading frames with some overlapping, see Figure 1. Gag-Pol is a polyprotein that encodes gag and pol in different reading frames. Gag is in the default reading frame. To synthesize pol, the HIV-1 RNA 3D structure sometimes causes a ribosomal frame-shift, i.e. changes the reading frame. Gag, pol and env are polyproteins that undergo cleavage too. Proteins tat and rev are synthesized by splicing exons [118]. In addition to this complexity there is an anti-sense protein (ASP) in HIV-1 that could be expressed in vivo [17], but is in the process of receiving complete acceptance by the HIV-1 research community, and for that reason is not depicted in Figure 1 [192].

For the HIV-1 genome there are 3D structural elements like the TTTTTT slippery site, followed by a stem-loop structure, that regulates the ribosomal frame-shift in gag-pol to change from the gag to the pol reading frame [156]. Another element is RRE (Rev response element) in the env coding region that comprises around 350 nucleotides and has a scaffold structure for interacting with Rev proteins.

3.4 PROTEOMES

Viral proteins are usually classified into structural (capsid, nucleocapsid), regulatory, enzymes and accessory (auxiliary). Focusing the discussion in retroviruses, their genomes encode three main groups of proteins: core, enzymes and envelope with genes called gag, pol and env respectively. Gag stands for *group specific antigens* or assemblin, pol for polimerase (enzymes) and env for envelope proteins. Other

proteins outside these genes are considered as regulatory, virulence factors or accessory. The virion 3D structure is related to core proteins encoded in gag as can be seen in the HIV-1 virion represented in Figure 2.

Although viral proteins might have many functions, in the next section we will mention some of them for HIV-1.

3.4.1 *HIV-1 proteins*

Gag is cleaved by the protease encoded in pol into structural proteins ma (matrix), ca (capsid), nc (nucleocapsid), and polypeptide p6. The structural elements are sufficient to form the viral particle by self-assembly. Gag proteins are responsible for membrane targeting and binding, Gag-Gag multimerization, packaging the viral genome as well as env proteins into virions, budding and release of virions [138].

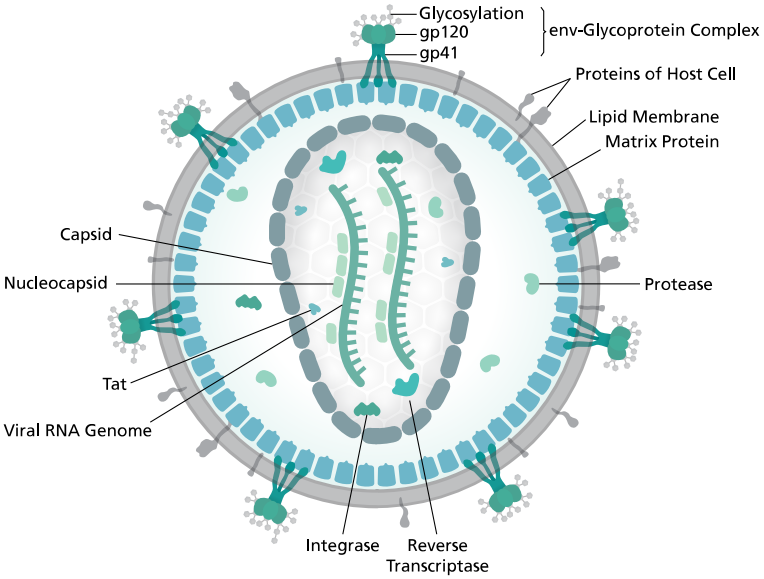


Figure 2: HIV-1 Proteome.
Cleavages: gag into {capsid, nucleocapsid, matrix, p6}, env into {gp120 (SU), gp41 (TM)}, pol into {protease, integrase, reverse transcriptase}. Copyright Wikipedia ©2016.

The pol polyprotein is cleaved into enzymes pr(protease), rt(reverse transcriptase) and in(integrase). Protease cleaves the viral polyproteins. Reverse transcriptase transcribes viral RNA into DNA and integrase integrates viral DNA into the human genome.

Envelope proteins are gp120 (SU) and gp41 (TM), responsible for attaching and fusing the virion to the cellular membrane. Protein gp120 binds to the CD4 protein in human T-cells [138]. The env (gp160) protein can be cleaved by human proteins furin and PC1 into gp120 and gp41 [48].

Rev and tat are regulatory proteins. Tat (transactivator) activates transcription initiation and elongation [30]. Rev is a phosphoprotein that binds to the viral mRNA structural element RRE and promotes nuclear export, stabilization, and utilization of the viral mRNAs containing the element.

Nef downregulates CD4 and MHC class I molecules, and alters intracellular traficking pathways [168]. Vpr (viral protein R) targets the nuclear import of preintegration complexes and causes cell growth arrest [121]. Vpu (viral protein U) degrades CD4 in the endoplasmic reticulum and enhances virion release from the plasma membrane [84].

Vif (viral infectivity factor) marks the human protein APOBEC3G for polyubiquitylation and proteasomal degradation hijacking the complex Cullin5 E3 ubiquitin ligase, composed by the proteins ElonginB, ElonginC, Cullin5, and Rbx2. The human protein APOBEC3G (apolipoprotein B editing catalytic subunit-like 3G) increase the mutation rate of viral genomes, interferes with reverse transcription and viral DNA integration [131, 166].

3.5 INFECTIOUS CYCLE

To replicate, viruses must enter into their host cells [182], move through them [202], evading the immune system [193], hijacking cell regulation [40] and subverting the host protein synthesis machinery [201]. Most viruses also need access to the cell nucleus [35] to achieve their replication.

Structurally, a virion has a short genome in RNA or DNA surrounded by a protein coat (capsid). Lacking metabolism and movement, viruses enter into animal hosts through the air or fluids and travel along mucus layers, the blood stream, motile cells and neuronal pathways until reaching cells of a type they can penetrate (susceptible cells). Viral proteins in the virion must execute the stages of infection of a host cell: attachment, penetration, replication, assembly and release to other cells.

At all stages, viral proteins must sense the cellular environment to detect cues, like changes in pH level, that trigger conformational changes. These changes allow viruses to achieve milestones like fusing with the host membrane to penetrate the host cell and uncoating its proteins and genome close to the nucleus. This is achieved by maintaining the viral proteins in a meta-stable state that can be easily relaxed without the need of an energy supply [182].

Viruses send signals to the host cell in order to induce changes in the membrane to facilitate their entry, neutralize immune system defenses, and hijack host proteins to perform specific tasks like moving towards the nucleus or towards the membrane to infect other cells

[40]. Intracellular transport is achieved by using endocytic vesicles as vehicles or by binding to and controlling human motor proteins.

The cue detection and the hijacking mechanisms of the virus are largely based on human-viral protein-protein interactions. Some of these interactions are based on structural or sequence similarity.

3.6 VIRAL BIOINFORMATICS

Curated viral genomes, protein sequences and alignments can be obtained from the NCBI viral genomes resource [23]. General VHPPIs are stored in databases VIP DB [32], VirHostNet [140] and VirusMINT [28]. There are specialized VHPPI databases for Hepatitis B [90] and Hepatitis C viruses [114].

3.6.1 *HIV-1 bioinformatics and systems biology data sources*

The reference genome for HIV-1 in the NCBI has the reference sequence code (RefSeq) NC_001802.1. The HIV-1 database for curated sequences and alignments used in this thesis is in the NIAID [111]. The NIAID also curates human-HIV-1 PPIs obtained from research papers in a database that is used in this work to validate predicted interactions [79, 2]. Overviews of the PPIs in the database are reported in references [162, 158].

The role of HIV-1 disordered protein content is reviewed in reference [211]. The human pathways affected by HIV-1 are analyzed in references [31, 165] and the perturbed subsystems in reference [127].

4

SYSTEM BIOLOGY AND BIOINFORMATICS OF VIRUS HOST INTERACTIONS

4.1 BACKGROUND

In the co-evolutionary arms race between humans and viruses it is important to understand the virus structure and function (see Chapter 3), the human susceptibility (population genetics), as well as the human-virus interactions.

Human susceptibility to infectious diseases can be investigated using genome wide association studies (GWAS) that use human genomes and clinical data as input. Single nucleotide polymorphisms (SNP) are used as markers in the genomes of big groups of patients. The SNPs inside the exon or regulatory region of a gene that are highly correlated with viral diseases symptoms suggest that the gene product encoded by the region that contains the SNP is targeted by viruses [103, 104]. The SNP also can be located in a regulatory, non-coding region. For instance, an SNP in the 5' flanking region of the MICA gene was associated with progression from chronic hepatitis C to virus-induced hepatocellular carcinoma using a sample of 721 Japaneses [113], another study found an association between the impaired immune response to Hepatitis B vaccination and a SNP in the 3' downstream region of FOXP1, a transcription factor involved in B-cell development, using a sample of 981 Indonesians [43].

A GWAS for viruses, like it has been conducted on bacteria, is more difficult to undertake because the mutation rate is huge, especially for RNA viruses [103].

As the interactions between the viral and human genome products like proteins and miRNA are essential to understand viral infections, the purpose of this chapter is to show how the bioinformatic and systems biology tools for virus-host interactions can help to integrate big amounts of information and guide in the formulation of new hypothesis. We focus on the more documented kind of interactions, protein-protein interactions.

4.2 PROTEIN STRUCTURE AND FUNCTION

Proteins are created as polymers, chains of amino acid residues linked sequentially. In many cases, as the chain grows, is folded or coiled into a stable 3D configuration that confers the protein a functional conformation, a biological mission inside the cell. Their structure is usually studied in four levels: 1) primary structure: the sequence of

amino acid residues, 2) secondary structure: small protein regions with a typical 3D structure like α helices and β strands, 3) tertiary structure: the coordinates for every atom in the protein and 4) quaternary structure: the 3D structure of proteins with multiple subunits.

The accumulated knowledge about protein functions can be found in the Gene Ontology (GO) database [7] in which there are expert annotations for proteins: the biological processes in which they participate, the cellular component they belong to, and the functions they perform.

4.2.1 *Domains and SLiMs*

In addition to the primary to quaternary structures referred above, protein structure is studied by finding reusable, modular protein portions with particular functions: domains and SLiMs. Finding clues to protein functions often begins by the identification of domains and SLiMs and this task begins by aligning protein sequences to find conserved regions.

Domains are defined in reference [159] as:

> spatially distinct structures, within proteins, that could conceivably fold and function in isolation.

Their size ranges from 25 to 500 amino acid residues and are cataloged in data bases like structural classification of proteins (SCOP [6]) and CATH [180].

SLiMs are small protein regions of 3 to 12 amino acid residues with a particular biological function with relative independence of the rest of the protein like signaling, post-traslational modification, binding or targeting [41], generally the function is fulfilled through the interaction with protein domains of other proteins [49]. They are used extensively by eukaryotes in cell signaling and regulation [49]. The database for SLiMs in eukaryotes is ELM, standing for the eukaryotic linear motifs resource for functional sites in proteins [51].

Usually, SLiMs are represented computationally as regular expressions. A SLiM instance is a protein subsequence that matches the regular expression. For instance, a SLiM represented by the regular expression **R.[RK]R.** have several instances like **RVRRE** in Ebola virus [199] and **RKRRF** in Human respiratory syncytial virus A2 [186].

4.2.2 *Protein disorder prediction*

Many proteins are intrinsically disordered, i.e. the proteins do not reach a stable or folded configuration. This gives them the flexibility of changing their conformation in several ways as they interact with other molecules [85]. Disorder is fundamental to protein function in two ways: 1) disordered regions often contain accessible post-

translational modification (PTM) sites, 2) disordered regions domains adopt different structures with different partners, increasing function repertoire without increasing genome size [14]. Disordered regions also facilitate the formation of complexes and assemblies and the mediation of regulated conformational changes [115].

The 3D structure for proteins is stored in the Protein Data Bank [15] while the information for intrinsically disordered proteins is in the DisProt database [179]. Many proteins have a mixed structure with ordered and disordered regions. The techniques for structure determination like X rays do not produce good results with protein disordered regions [85].

Uversky et al. realize that evolution pushes viruses to adapt to high mutation rates and not too high thermodynamic stability [210]. For this reason, viral proteins have a low number of inter-residue interactions and a high occurrence of polar residues that account for the abundance of disordered content.

A recent review on bioinformatic methods for protein disorder predictions highlights the complexity of the task [57]. There are several definitions of disorder and each method performance is good for the definition it is based on. In general, the better methods reach a 70% of prediction accuracy of disordered residues in the CASP experiments conducted by the Protein Structure Prediction Center, and a 10% missclassification rate for ordered residues. The techniques for disorder prediction take as input a protein sequence and return a value, usually between 0 and 1 for every residue, quantifying is disorder level. With these values, disordered regions can be marked.

The methods can be classified in two categories: machine learning and physical. The machine learning methods train classifiers like neural networks and support vector machines with data sets mostly obtained from Disprot [179] or proteins with missing residues in their X-ray structures obtained from the PDB [179]. As Disprot data data is scarce and the PDB is biased to globular proteins, all these methods are impaired by lack of data.

4.2.2.1 *IUPred*

This is a physical method based on two observations on the proteins with disordered content: i) they are depleted of typically buried residues and this can be measured by low mean hydrophobicity, ii) they are enriched with typically exposed residues and this can be measured by high net charge. Both observations can be summarized in the rule:

> If a residue is not able to form enough favorable intrachain contacts it will not adopt a stable 3D position [56].

From this point of view, globular proteins have sequences that favor a good number of intrachain contacts and disordered proteins do

not. Inter-residue interactions are modeled by force fields (statistical potentials or energy functions) that have been used successfully in fold recognition, docking and predicting protein stability.

The main hypothesis is that the primary structure of a globular protein determines its total energy, and this is the lowest level attainable by the sequence at the optimum level of inter-residue interactions. The IUPred algorithm establishes the total energy of a protein as the sum of the statistical potentials:

$$E = \sum_{ij=1}^{20} M_{ij} C_{ij} \tag{1}$$

- where C_{ij} is the number of interactions between residues of types i and j
- and M_{ij} is the interactions energy between residues of types i and j

The total energy per amino acid is the sum of the statistical potentials

$$\frac{E_{est}}{L} = \sum_{ij=1}^{20} n_i P_{ij} n_j \tag{2}$$

where:

- L is the length of the primary sequence
- $n_i = N_i/L$ is the frequency of residues of type i (N_i is the number)
- and P_{ij} quantifies the pairwise energy between amino acids i and j

If many globular proteins from the Protein Data Bank are considered, the total energy of the kth protein is

$$E^k = \sum e_i^k \tag{3}$$

where e_i^k is the energy of all residues of type i interacting with the rest

This energies are calculated by equation (1):

$$e_i^k = \sum_{ij=1}^{20} M_{ij} C_{ij}^k \tag{4}$$

The approximation to these energies, using equation (2) is

$$e_i^k(\text{estimated}) = N_i^k \sum_{j=1}^{20} P_{ij} n_j^k \tag{5}$$

The IUPred algorithm is based in finding the ith row of the matrix $\mathbf{P}$, given by the minimization of:

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k)^2$$

making $\delta Z_i / \delta P_{ij} = 0$ for all P_{ij}

One of the results of the IUPred paper is the matrix $\mathbf{P}$ [56]. The authors conclude that some of the eigenvalues of matrix $\mathbf{P}$ are representing: hydrophobicity, cysteine abundance, structure breaking amino acids (proline, asparagine, glycine) and net charge of the protein. IUPred output for each residue in the sequence is a disorder value between 0 and 1. Residues with values higher that 0.5 can be considered unstructured.

Hagai et al. found that the use of a window of 10 residues improves the predictions of IUPred [86]. They compute the IUPred disorder for every amino acid residue in a protein. Then, they compute for every residue the average disorder value of a window of 10 residues centered in the residue. Regions of the protein with averaged disorder values greater than 0.4 are considered as disordered.

4.2.3 *Relation between disorder and SLiMs*

SLiMs occur more frequently in viral protein disordered regions [80], in different amounts between viral families [163]. Viral hubs, proteins that have many interactions with host proteins, tend to have more disordered regions [133].

Viruses use mimic host SLiMs as interaction sites with host proteins in multiple ways [98]. For example to interfere with host signaling proteins [3].

The presence of SLiMs in eukaryotic viruses is prevalent [87]. There is a correlation between the amount of disordered content and number of linear motifs in viral proteins [87]. It seems that disordered regions in viral proteins, free from evolutionary pressure to high thermodynamic stability, can evolve quickly short linear motifs de novo to mimic the host in order to create new interactions [42]. As eukaryotic hosts use linear motifs extensively for cell regulation [196], the referred evolution of new viral motifs can be an important form of attacking the host.

The abundance of disordered content in viral proteins can be interpreted as a weapon, to evolve new SLiMs to wire new interactions, and as an armor, a defense mechanism the virus can use to disrupt SLiM-mediated interactions that normally take place in the host cell in order to evade the immune system or avoid apoptosis.

4.3 PROTEIN-PROTEIN INTERACTIONS

Protein-protein interactions (PPI), are defined in reference [47] as:

> specific physical contacts between protein pairs that occur by selective molecular docking in a particular biological context.

Many of the protein functions imply PPIs [47]. These interactions are mediated by electrostatic interactions, hydrogen bonds, van der Waals forces and hydrophobic effects. It is known that the alteration in common protein-protein interactions is related to diseases (see [97], chapter 9).

Although difficult to validate [126], there are established experimental methods for detecting PPI [9, 157]. The most used ones are the binary yeast two hybrid assay (YTH) [187] and the two co-complex techniques: tandem affinity purification coupled to mass spectrometry (TAP-MS) [36] and coimmunoprecipitation (CoIP) [19]. Other techniques are reviewed in reference [21].

As there has been small overlap between the PPI data obtained with different techniques [47], a method for assigning confidence scores to PPI was proposed to alleviate this situation [22]. It is based on combining several experimental methods and a strict logistic regression model to compute a confidence score for any candidate PPI. The rationale for this score is to use it to exclude false positives.

Structurally, PPI might take place between domains of the proteins or between a domain and a SLiM. SLiMs are preferentially found in protein disordered regions and are relatively conserved [41].

PPI allow proteins to form groups in order to fulfill important functions and the groups allow us to analyze biological processes from a higher level point of view. The most used protein group concepts are network motifs, protein complexes, protein functional modules, and pathways. See section 4.4 below.

4.3.1 *Pathogen-host interactions*

The pathogen-host (protein-protein) interactions (PHI) serve as bridge between the virus and the host biology through an important subset of host proteins: the pathogen targeted ones. These proteins are the point of departure in a systems biology analysis of virus-host interactions.

When we know that a viral protein interacts with host proteins, this information can be contextualized in the the host protein-protein interaction network with the functional and hierarchical information annotated using the protein group concepts (see section 4.4 below). This allows to postulate pathogen strategies for infecting cells.

There are several ongoing efforts for cataloging and curating PPI which have produced several databases, summarized in Table 2. BIND

catalogs interactions between proteins, RNA, DNA, molecular complexes, small molecules, photons and genes for many species of organisms [8], BioGRID focuses on model organisms [29], DIP [172], IntAct [100] and MINT [122] only register PPIs. HPRD only considers human PPI [101]. APIDB [160] and PINA [37] aim at integrating data from several databases.

For PHI there exist the databases HPIDB [112], MPIDB [83], Patric [82], PHI-base [206] and PHISTO [61]. For virus-host interactions: VIP DB [32], VirHostNet [140] and VirusMINT [28]. Also there are databases for interactions between humans and specific viruses like HIV1 [79], Hepatitis B [90] and Hepatitis C [114].

The diversity and quantity of PPI databases causes problems like incompatibility of formats and duplication of curation efforts [148] that integration initiatives like the IMEx consortium are trying to solve [152].

Another concern with the PPI data sets is the incorporation of protein structure at the level of domains, SLiMs, and more comprehensively, full three-dimensional structure. The scarcity of data in the protein data bank (PDB) is the obstacle for this [15]. However, data is increasing and there is a database specialized in protein-protein interactions with known structural information, and even tools to predict interactions of the kind SLiM-domain: the database of three-dimensional interacting domains (3did) [137].

4.3.2 *SLiM mediated interactions*

Viruses tend to interact with universally expressed, functionally diverse and slow-evolving human proteins. These virus-targeted proteins are enriched in disordered regions, SLiMs and binding interfaces [88]. SLiMs are in signaling pathways proteins like kinases and participate in cell cycle regulation, protein degradation and proteolytic cleavage.

If VHPPIs are classified into domain-SLiM and domain-domain, there are differences in them. Domain-SLiM interactions tend to be located in the cytoplasm, the mitochondria and the nucleus while domain-domain interactions tend to be in extracellular space, cellular membrane, cytoskeleton and the peroxisome [88].

Viral SLiM mimicry makes domain-SLiM VHPPIs more predominant than domain-domain interactions. Furthermore, domain-SLiM interactions are used by viruses that belong to different families, suggesting convergent evolution of mimicked SLiMs as a mechanism to interact with human systems [88].

With over a million estimated SLiMs in the human genome [191], there is a call to develop future experimental methods to detect SLiM-mediated interactions [18]. Meanwhile, the development of bioinfor-

Table 2: Protein protein interaction databases

Database	**URL**	**Ref.**
BIND	http://bond.unleashedinformatics.com/	[8]
BioGRID	http://thebiogrid.org/	[29]
DIP	http://dip.doe-mbi.ucla.edu/	[172]
HPRD	http://www.hprd.org/	[101]
IntAct	http://www.ebi.ac.uk/intact/	[100]
MINT	http://mint.bio.uniroma2.it/mint/	[122]
MIPS-MPPI	http://mips.helmholtz-muenchen.de/proj/ppi/	[154]
APIDB	http://eupathdb.org/eupathdb/	[160]
PINA	http://cbg.garvan.unsw.edu.au/pina/	[37]
3DID	http://3did.irbbarcelona.org/	[137]
PHI databases		
HBVdb	http://hbvdb.ibcp.fr/HBVdb/	[90]
HCVpro	http://cbrc.kaust.edu.sa/hcvpro/	[114]
HPIDB	http://www.agbase.msstate.edu/hpi/main.html	[112]
HIV1-Human	http://www.ncbi.nlm.nih.gov/projects/RefSeq/HIVInteractions/	[79]
MPIDB	http://jcvi.org/mpidb/	[83]
Patric	http://patric.vbi.vt.edu/	[82]
PHI-base	http://www.phi-base.org/	[206]
PHISTO	http://www.phisto.org/	[61]
VIP DB	http://vipdb.cgu.edu.tw/	[32]
VirHostNet	http://pbildb1.univ-lyon1.fr/virhostnet/	[140]
VirusMINT	http://mint.bio.uniroma2.it/virusmint/	[28]

matic methods for inferring SLiM-mediated interactions can give interesting candidate interactions.

4.4 GROUPS OF PROTEINS

4.4.1 *Network motif*

A network motif tries to capture the notion of recurring, significant patterns of interconnections in a network. The first definition for network motifs was given as particular subgraphs occurring in a network a number of times significantly higher than expected [134]. The expected number of subgraph occurrences in a network is computed from a suitable network model, free scale networks for example [10], and compared with the actual found number. If the analytical calculation is too difficult an alternative approach consists in generating, randomly, a large set of networks with the creation algorithm of the network model, and using it as a sample to estimate the expected number of occurrences of a particular subgraph with direct counting.

Some network motifs have been studied as blocks or circuits with specific functions that embody solutions to recurrent "design" problems, like the feed-forward loop motif. This motif is an abstraction, in a directed network, of two inducers, two transcription factors and a promoter acting in a specific topology to regulate a gene [129].

In this sense, network motifs are the first level of complexity to study in a biological system and there are recent attempts to explain how these motifs might evolve in model gene regulatory networks [25].

4.4.2 *Biological network motif*

Wong et al. take the notion of motif as a subnetwork or subgraph over represented in a network and expand it with the application-dependent requirement of biological significance, giving the notion of *biological network motif* [106]. They also relax the network motif definition by not requiring to find the same link structure in the subgraphs, only demanding that the subgraph be connected.

The authors leave room for different ways to measure the biological significance of a subgraph and propose three concrete ways to do it, developing five different algorithms based on them. The three measures of biological significance are called motifs included in complex, motifs included in functional module and GO term clustering score. The first two check if the network motif is part, i.e. is a subgraph, of a protein complex or functional module; the last one uses a clustering score for a subgraph based on the GO database that assigns functions to proteins [7].

4.4.3 *Modules, complexes and pathways*

Node or vertex communities are a second level of complexity and the idea is to find algorithmically "natural" clusters or groups of vertices interacting more strongly between them that with the rest of the network [147]. Like the different clustering problems in computer science, finding network communities is computationally hard and there are several methods proposed to do it [75]. Some are based on the *modularity* metric of a network [146], and others are based on the betweenness of edges [34].

In systems biology the most important applications of the methods for finding communities are the discovery of potential regulatory and signaling *circuits* in networks [91] and the identification of candidate protein complexes and protein functional modules [183].

Protein complexes and functional modules have a more biological definition than network motifs. Usually, they group more proteins than network motifs, but have an important difference about the timing and location of the PPI. In a protein complex all the proteins interact at the same time in a definite cellular place to form super-structures like the spliceosome [213] while a functional module groups proteins that work at different times and places in order to fulfill a specific function like controlling the cell cycle [183].

A pathway has a wider functional definition, it comprises many molecules that, acting together, lead to a significant cellular change or product, for example the glycolysis pathway. Pathways can be classified as regulatory, signaling or metabolic and can contain several protein complexes, functional modules and network motifs. The bigger databases for pathways are KEGG [99] and MetaCyc [27], compared in reference [4].

With the protein group concepts introduced above, the catalog of endogenous PPI in an organism allow us to understand its functional and hierarchical organization from a systems biology viewpoint.

4.4.4 *Viral contacts with groups of proteins*

Viruses interact with "clubs" of strongly interrelated proteins [209]. For this reason the notions of protein modules, protein complexes, pathways and network motifs can be useful to give meaning to a viral-human protein-protein interaction. This contextualization can be done using tools like gene ontology [130] and KEGG pathways [175].

4.5 PROTEIN INTERACTION NETWORKS

Protein interaction networks (PINs) are used to represent and show graphically the interactions between proteins. Nodes are proteins and

links represent interactions between pair of proteins. Endogenous PIN help us to understand the cooperation that takes place between proteins in an organism to achieve biological functions.

For example the Figure 3 shows the endogenous PIN for the Epstein-Bar virus (EBV) obtained with the Y2H method. The authors did an evolutionary analysis grouping EBV proteins in herpes viruses core (yellow) and non-core (green) and found that they tend to interact more with other proteins in the same group [26].

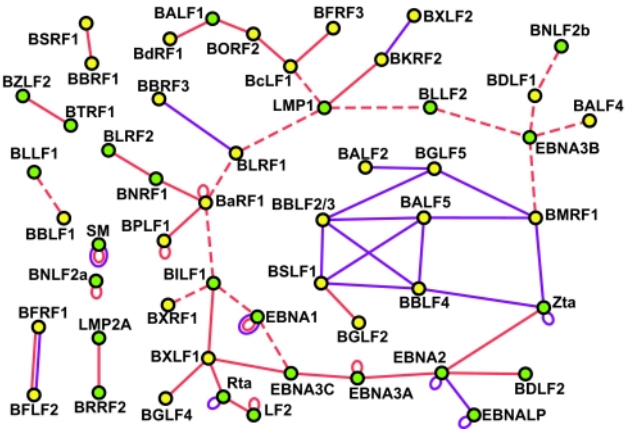


Figure 3: Endogenous Protein Interaction Network for EBV. Proteins are divided in two groups: herpes viruses core (in yellow) and non-core (green). Purple arcs represent interactions known before the experiments and pink arcs represent the interactions found. Low confidence interactions are represented with dashed lines. Copyright © Proc Natl Acad Sci USA. 2007 May 1; 104(18): 7606–7611. doi: 10.1073/pnas.0702332104

PHI networks present us with the pathogen and host (targeted) proteins as nodes. Every link connects a pair of proteins, one of the pathogen and one of the host. Fore example, the Figure 4 show the PHI network for EBV and human proteins.

In the Table 3 there is a compendium of papers with endogenous PPI and PHI networks for viruses that affect humans. For the Hepatitis C virus there are several works: one based on the Y2H method [73], other focusing on nonstructural proteins[50], other based on genome-wide small interfering RNA (siRNA) screen using an infectious HCV cell culture system [120] and other using Y2H [44], increasing the picture of the interactions with the virus. The network for interactions with Dengue virus began was reported in reference [102] and expanded in reference [55] with new interactions.

4.6 NETWORK THEORY ANALYSIS

The most used network metrics are the degree and betweenness centralities. Degree is the number of links a node has, and centrality

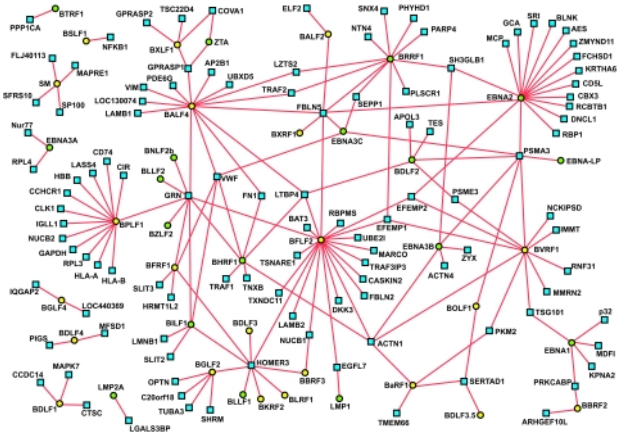


Figure 4: Pathogen-Host Interaction Network for human and EBV Human proteins are represented as cyan squares. EBV proteins are circles, yellow fore core herpes viruses proteins and green for non-core. Copyright © Proc Natl Acad Sci USA. 2007 May 1; 104(18): 7606–7611. doi: 10.1073/pnas.0702332104

is the ratio between the number of minimal paths that pass through a node and the total number of minimal paths between all pairs of nodes. The degree distribution has the proportion of nodes for every possible degree, and can also be interpreted as the probability of choosing a node of a given degree at random. Central nodes measured by degree are called hubs and measured by betweenness, bottlenecks. Another useful kind of node is the bridging node, a node representing a protein that belongs to two or more protein modules.

One way to measure hierarchical organization is given by the local clustering coefficient. This is defined for a node as the ratio of the number of links between the node and all its neighbors and the maximum possible number of links between them.

Another set of metrics related to network organization consider assortativity or assortative mixing, the tendency of similar nodes to share links. One of them is the assortativity coefficient, in which the similarity is considered as the similarity of the nodes degrees. If the nodes in a network with similar degree do not tend to link between them, the network is called diassortative.

A network model comprises several topological properties of a network like the degree distribution and an algorithm to create networks with these properties. The properties allow us to reason about the system modeled with the network and the algorithm allow us to create many instances of networks with the same properties in order to do statistical analysis and simulations. For instance, a network created with an small-world network algorithm will have a small average distance between the nodes [5].

One model used to reason about PPIs is the sticky network [161] which is based in giving a stickiness index to each protein based on

Table 3: Publications with virus-human PPI networks

Virus	**References**
Arena viruses	[58]
Chikungunya virus	[20]
Coronaviruses, SARS	[67],[200], [155]
Dengue virus	[102],[128]
Epstein-Barr virus	[26], [76],[92]
Hepatitis B virus	[119],[90]
Hepatitis C virus	[73],[50],[44],[120],[114]
Human immunodeficency virus 1 (HIV-1)	[93],[94],[79]
Human Respiratory Syncytial Virus	[207]
Herpes simplex virus type 1 (HSV-1)	[198],[116], [76],[92]
Human T-lymphotropic virus 1 (HTLV-1)	[181]
Human T-lymphotropic virus 2 (HTLV-2)	[181]
Influenza viruses	[178],[203],[46]
Kaposi's sarcoma-associated herpes virus (KSHV)	[194], [170],[76], [92]
Poxviruses, Vaccinia	[197], [212]
Vesicular stomatitis virus (VSV)	[135]
Varicella-zoster virus (VZV)	[194], [76],[184]

the number of its interacting domains. The probability of an interaction between two proteins is computed as the normalized product of the two proteins stickiness indexes. Other common network model is the scale-free [10] one.

4.6.1 *Human protein interaction network*

The human PIN or interactome has been in construction since the completion of human genome project [117] and has been expanded considerably [185, 171]. It has been analyzed in comparison to some model organisms [81] and a recent analysis shows that the core of the network has been mostly revealed [107].

The partial data suggests that small world and scale-free network models capture some of its topological properties: short average distance between nodes [204], closeness to a power-law degree distribution [96], hierarchical organization inferred from node clustering, diassortativity by degree [145] and correlation between biological importance of nodes and their measured degree and betweenness centralities [95].

4.6.2 *Intraviral Protein Interaction Networks*

The intraviral or endogenous PIN of a subset of the viruses in Table 3 were analyzed in reference [133]. The authors concluded that free-scale and small world network models do not fit well to these PINs, the best fitting model is the sticky network model but it does not apply to all the viruses analyzed [161]. The networks are diassortative by degree and network resilience is good if random and deliberate node removal is simulated on them.

4.6.3 *Pathogen-Host Interaction networks*

In networks of interactions between host and pathogen proteins it is convenient to discriminate the human and viral hubs. In this context, a human hub is a protein that interacts with many viral proteins and a viral hub is a viral protein that interacts with many human proteins. With the available data viruses, the authors of references [133],[64], and [176] conclude, in agreement, that virus proteins target human hubs and bottlenecks.

Dyer et al. found that viruses disrupt the same human cellular processes, i.e. control of the cell cycle, apoptosis, immune response and transport of molecules through the nuclear membrane, even though they target different proteins [64]. They also found that human targeted proteins are often located in cancer pathways, a correlation that is reviewed in [69].

Meyniel-Schicklin et al. found that the majority of viral proteins target a small number of human proteins and that viral hubs are a minority [133]. They compared the predicted structural disorder in viral hubs and the rest of viral proteins concluding that the first ones tend to have more structural disordered regions, as predicted by the tool DisEMBL [123]. This is in line with the finding that viral proteins tend to accommodate- SLiMs in their disordered regions [80], in this way a viral protein with more disordered regions would tend to have more- SLiMs to interact with human proteins.

Another study incorporating protein structure information found that viral proteins tend to interact with human proteins using interfaces that participate in human endogenous interactions by binding, transiently, to multiple human regulators [77]. Wuchty et al. find that HIV-1 proteins target groups, called by them “clubs”, of human proteins strongly intertwined [209]. Navratil et al. use the diseosome, a network that links diseases with proteins, to explore correlations between virus targets and diseases, and conclude that viruses also target bridging proteins, proteins that connect different protein modules [142].

Part II

MATERIALS AND METHODS

5

CALIBRATING THE METHOD OF SLIM-MEDIATED VHPPI PREDICTION WITH HIV-1 DATA

5.1 INTRODUCTION

In order to predict SLiM-mediated VHPPIs we implement and compare the strategies for SLiM filtering: conservation above a threshold of the available viral sequences, localization in a protein disordered region and rarity, or difficulty to form by pure chance. Each filtering method produces a set of SLiMs – conserved (C), disordered (D) and rare (R). With sets C, D, R we form derived union and intersection sets. Each of these sets allow us to predict interactions between the viral protein containing the SLiM and the host proteins that interact with the SLiM.

All the sets generated are compared by filtering strength. They also are compared by the number of PPIs derived from the set that have supporting evidence in a database –i.e. correctly predicted. The comparison by number of the PPIs correctly predicted by set allow us to rank the PPIs partially.

To conduct the comparison of the sets we choose HIV-1. It is the virus with more bioinformatic data available, with the NIAID databases for sequences and alignments [111] and for interactions with human proteins [79]. We also use the HIV-1-human PPIs mediated by SLiMs as reported in the LMPID database [174].

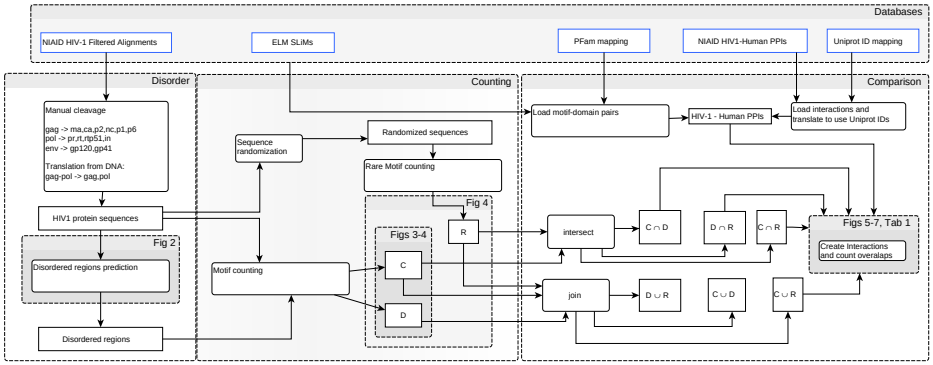


Figure 5: Methods for VHPPI prediction.

We divide the calibration of the VHPPI prediction method in three parts: 1) preprocessing and disordered regions prediction, 2) counting of SLiMs, and 3) comparison of interactions against NIAID and LMPID databases.

5.2 PROTEIN SEQUENCES AND PREPROCESSING

We download alignments for HIV-1 proteins env, gag, nef, pol, rev, vif, vpr, tat, vpu for the year 2014 and an alignment of Gag-Pol DNA sequences with years previous to 2015 from the NIAID HIV-1 sequence database [111]. Gag-Pol sequences were translated following reference [156]: of 3648 sequences, 3626 containing the slippery sub-sequence TTTTTTA were used to perform a computational translation considering the frame-shift at the given sub-sequence, Section A.3.

We filter all protein sequences by HIV-1 sub-types B and C for their worldwide dominance and computationally cleave some of the alignments in the following manner: env into gp120, gp41, pol into pr, rt, rtp51, in and gag into ma, ca, p2, nc, p1, p6 [79]. After the cleavage we eliminate the gaps and asterisks in the resulting alignments in order to reinterpret the files as sets of sequences, Figure 5, Disorder panel at the left.

5.2.1 *Protein Disorder Prediction with IUPred*

Among several disorder prediction algorithms for proteins [57] we use IUPred [56]. This predictor implements a physical model based on force fields between residues statistically calibrated with a set of globular proteins in PDB [56]. Its performance is comparable to other predictors [166] and can be installed locally.

IUPred is enhanced with a sliding window addition proposed by Hagai et al. that allows to define disordered regions [86]. Residues with IUPred computed values higher that 0.4 are considered disordered. For each residue an average disorder value is computed considering the IUPred values for surrounding residues in a window of size 10. This averaging is justified because the disorder tendency of the neighbors of a residue influence its disorder tendency. Residue windows with average disorder value higher than 0.4 are considered as disordered.

As IUPred receives as input a Fasta file with only one sequence, we split Fasta files with multiple sequences, call IUPred on every split sequence-file, compute the sliding window based average values and give as output a list of disordered regions per protein sequence id. We set the parameter *long* when calling IUPred, see Figure 5, Disorder panel at the left.

5.3 SLIM COUNTING

We download all the SLiMs, instances and interactions from the ELM database [52] and create an in-memory ELM data structure with each SLiM identifier, its regular expression, its instances and its interac-

tions with protein domains. We write scripts to compute: the number of sequences with a given SLiM, the number of SLiM instances per protein, the number of SLiMs conserved above a percentage of sequences (set C) and the number of SLiMs in disordered regions (set D).

We randomize the HIV-1 proteins to create a big data set. For each sequence in a protein file we create 1000 shuffled versions randomizing the residues located in disordered regions of the sequence, as computed with IUPred. Then, we count the rare (scarce) SLiMs in these shuffled data set, i.e. the SLiMs that are found in 1% of the randomized sequences or less (set R).

Based on C, D, R we create the union sets $C \cup D$, $C \cup R$, $D \cup R$, $C \cup D \cup R$ and intersection sets $C \cap D$, $C \cap R$, $D \cap R$, $C \cap R \cap D$. See Figure 5, panels Counting and Comparison (center at right).

5.4 PREDICTION OF PROTEIN-PROTEIN INTERACTIONS

We download the NIAID human–HIV-1 PPI database [79]. As the proteins in the database are identified by RefSeq records and the SLiM-domain interactions given by the ELM database are given by UniProt records, we map RefSeq to UniProt identifiers for human proteins using UniProt id mappings. We also download the LMPID database that curates virus-host ELM-mediated interactions [174].

For each motif set ($C, D, R, ...$) obtained per HIV-1 protein we create virus-human PPIs with base on the ELM database interactions and interacting domains. For each interaction reported in ELM we add the human protein interacting with the SLiM located in the viral protein. We also add the proteins that contain the domains listed as interacting with the SLiM. To map domains to human proteins we used the domain-protein mapping for the human proteome in the Pfam ftp server [195].

We create PPI networks from the sets of SLiMs obtained with the three filtering methods using the networkx python library [144]. See Figure 5, Comparison panel at the right.

5.5 COMPARISON OF FILTERING METHODS

To validate a prediction we use two sets: the NIAID HIV-1-human interactions and the set of ELM mediated HIV-1-human interactions, as identified in LMPID [174]. We count the number of hits, when a human protein predicted to interact with HIV-1 through one of the SLiM sets match a human protein interaction with HIV-1 in the corresponding validation set.

For all the SLiM sets obtained, and all the HIV-1 proteins, we analyze the overlap between the set of predicted human proteins interacting with HIV-1 and the set human proteins in NIAID interactions.

7. The null hypothesis is that the performance of our method is equal or lower than random sampling. Taking as universe all the human proteins, estimated as $n = 30057$ from reference [105], the set A of human proteins interacting with HIV-1 from the NIAID database, the set B of predicted proteins with our method, we are interested in the number $t = |A \cap B|$, the correctly predicted proteins interacting with HIV-1. The null hypothesis is that $t \leq |B'|$, where B' is a random sample with the same size as B.

If $|A| = a$, $|B| = b$ and n is the size of the universe set, the hypergeometric distribution gives the probability that a random variable X for the size of the intersection takes a value less than t:

$$Pr(X < t) = \sum_{i=1}^{t-1} \frac{\binom{a}{i}\binom{n-a}{b-i}}{\binom{n}{b}}$$

The p-value for hypothesis testing is the probability of $X \geq t$ given by $1 - Pr(X < t)$, our test is one-sided.

We iterate the hypothesis testing across two dimensions: one for the HIV-1 proteins and other for the SLiM set used to infer the protein-protein interactions. The Table 6 has the sizes for the PPI sets and the Table 7 has the p-values.

6

VIRAL SUBVERSION OF THE HOST PROTEIN-SYNTHESIS MACHINERY

6.1 INTRODUCTION

There is no known virus that encodes a complete protein-synthesis system. This implies that viruses are forced to use the HPSM to translate their mRNA into products: miRNA, peptides and proteins. Viruses must control the HPSM and disrupt innate host defense systems capable of disabling protein synthesis [201].

The control and disruption of host signaling pathways is conducted through VHPPIs like the ones DNA viruses engage with the PI3K-Akt-mTOR pathway (phosphatidylinositol 3-kinase-Akt-mammalian target of rapamycin) [24]. The consequences of virus-host PPIs can be as significant as the shutdown of host protein synthesis done by rotavirus protein NSP3 [153].

6.2 SEQUENCES AND DISORDER PREDICTION

We consider the HPSM proteins listed in reference [201] and mapped them to Uniprot identifiers in order to match protein entries in the ELM database [12].

We select viruses for their availability of protein sequences in the NCBI viral genomes resource: Influenza, Dengue, West Nile, Middle East respiratory syndrome coronavirus (MERS), Ebolavirus, Rotavirus and Zika [23]. For influenza we choose type A, subtype H1N1, for Dengue we choose type 1, for Ebola the Zaire species.

We download every viral protein for each virus. For all viruses we set the parameter region as any, the parameter Full-length sequences only and the parameter host as human. For Influenza AH1N1 proteins we set the parameters collapsed sequences and full-length only, with the exception of M1,M2 and NS2. The details for the viral proteins are summarized in Table 4 for virus Dengue-1 and for virus influenza A H1N1 in Table 5.

For viruses Dengue type 1, West Nile and Zika the NCBI viral genomes resource stores the complete polyprotein sequence that must be manually cleaved. The viral genomes stored in Genbank files are computationally translated to protein sequences that are used as reference for cleaving the polyprotein into viral proteins. HIV-1 gag-pol in HIV-1 that has a frame-shift is computationally translated with a particular script, see the Appendix A, Section A.3.

Table 4: Dengue 1 proteins

Protein	Size	Number of sequences
C-anchored	114	1525
C	100	1525
E	495	1525
M-precursor	166	1525
M	76	1525
NS1	352	1525
NS2A	218	1525
NS2B	130	1525
NS3	619	1525
NS4A	127	1525
NS4B	249	1525
NS5	900	1525
P2K	23	1525
pr	91	1525

The size of each protein is given in number of residues. As the polyprotein was computationally cleaved into the products, the number of sequences is the same for all.

Table 5: Influenza AH1N1 proteins

Protein	Size	Number of sequences
HA	575	5603
M1	272	9989
M2	97	9300
NA	470	3391
NP	499	898
NS1	237	1377
NS2	121	7290
PA	717	1890
PA-X	252	273
PB1-F2	101	225
PB1	759	1825
PB2	764	1988

The size of each protein is given in number of residues.

The disorder prediction for viral proteins is computed in the same way that was performed with HIV-1 for the method calibration.

6.3 CREATION OF SLIM SETS

We use the scripts developed for HIV-1 to compute the number of sequences with a given SLiM, the number of SLiM instances per protein, the number of SLiMs conserved above a percentage of sequences (set C) and the number of SLiMs in disordered regions (set D).

We randomize the viral sequences. For each sequence in a protein file we create 1000 shuffled versions randomizing the residues located in disordered regions of the sequence, as computed with IUPred. Then, we counted the rare (scarce) SLiMs in these shuffled data sets, i.e. the SLiMs that are found in 1% of the randomized sequences or less (set R).

With the sets generated we create the union set $C \cup D \cup R$ to infer interactions from it.

6.4 PREDICTION OF INTERACTIONS

We compute the SLiM instances in viral proteins for all the human SLiM regular expressions in the set $C \cup D \cup R$. With the SLiM instances we infer PPIs between humans and the corresponding virus using the SLiM-domain associations in the ELM database and the protein-domain associations in the Pfam database [72].

The interactions are filtered for the HPSM proteins listed in reference [201] and in the Ribosomal Protein Gene database (RPGdb), selecting the cytoplasmic ribosomal proteins for homo sapiens [139].

6.5 ANALYSIS OF THE INTERACTIONS

The PPIs inferred are analyzed statistically. The proteins in the HPSM are sorted by the number of interactions predicted with viral proteins. The viral proteins are classified by the number of interactions with different human proteins.

We classify the interactions as tentatively disrupting or bridging the human protein-protein interaction network. A viral protein that interacts with only one protein in the HPSM probably disrupts a pathway, while a viral protein that interacts with two or more HPSM proteins probably wires a new path.

Part III

RESULTS AND DISCUSSION

7

SLIM-MEDIATED VHPPI PREDICTION METHOD

7.1 A GENERAL METHOD TO IDENTIFY SLIM-MEDIATED VHPPIS IN EUKARYOTES

As SLiMs are computationally represented by regular expressions there is always a possibility of finding instances in viral sequences by pure chance. For this reason is important to develop SLiM filtering methods.

Three filtering methods are implemented and systematically compared: conservation, location in disordered regions and rarity. The combination of filters produces a method to predict virus-host PPIs and rank them. The comparison of filtering methods performance is conducted with the virus with more abundant data, HIV-1.

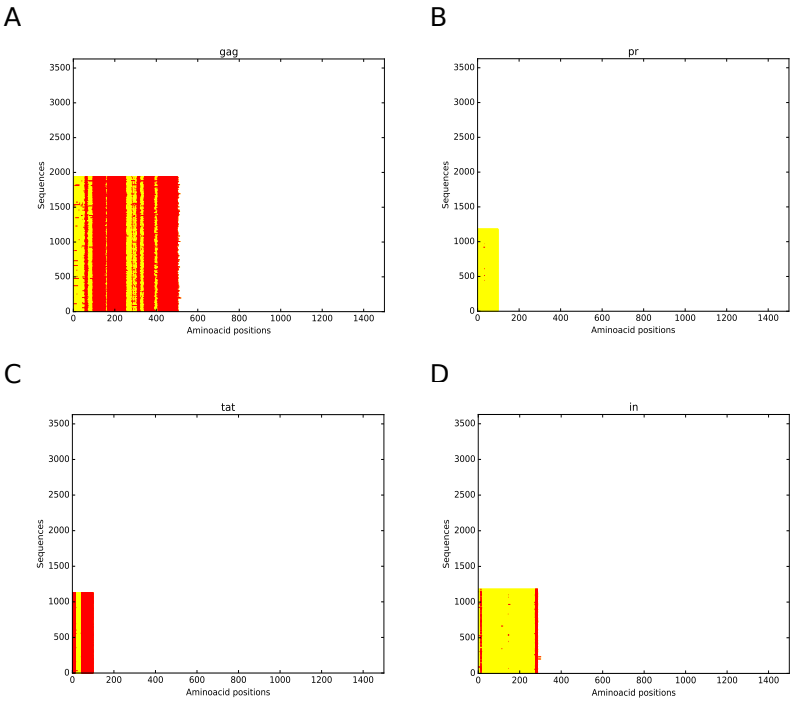
The developed method only use protein sequences as input and do not depend on protein 3D structures, for these reason it can be used with any sequenced eukaryotic virus to generate candidate PPIs. The restriction to eukaryotic viruses is based on the higher number of SLiMs in this kind of viruses and the use of the ELM database, because the ELM SLiM classes MOD (post-translational modification) and TRG (targeting sites) are less used in prokaryotes [87].

7.2 DISORDERED REGIONS AND SLIMS IN HIV-1 PROTEINS

We find that disordered regions for HIV-1 proteins are relatively conserved, Figure 6. Perhaps the viruses must keep flexibility in their proteins in order to interact with several partners.

Also, most of the conserved SLiMs in HIV-1 are located in protein disordered regions, Figure 7. A similar tendency is reported for the SLiMs that bind to SH2, SH3 and Ser/Thr Kinase domains [167].

The proteins that deviate the most from this tendency are vpr, vpu, gp41, in, and pr, with a percentage of conserved motifs that are located in disordered regions of 53.3%, 52.9%, 48.5%, 32.5%, and 0% respectively, Figure 7. The reason for this discrepancy lies in the few disordered regions predicted in the five proteins. Indeed, pr, in and gp41 are considered mostly ordered, while vpr and vpu are considered moderately disordered [211].



7

Figure 6: Disordered regions for HIV-1 proteins.
Each yellow line in the plots represents a protein sequence. The red segments denote disordered regions as deduced with IUPred with the sliding window addition explained in the Methods section. In Subfigures A and C the disordered regions for proteins *gag* and *tat* are displayed. In Subfigures B and D notice the lack of disordered regions for proteins *pr* (protease) and *in* (integrase), that explains the little overlap between disordered and conserved SLiMs in Figure 7.

7.3 ANALYSIS OF SLIM SETS OBTAINED

7.3.1 *A ranking of SLiM sets by filtering strength*

Considering the sizes of SLiM sets we can rank them by filtering strength, from the less filtering to the most. The obtained ranking is $R, D, C, C \cap D, D \cap R$. The criterion that that filters the most is location in a disordered region and rarity. It is followed by location in a disordered region and conservation,- Figure 8.

The sets $D \cap R$ (SLiMs hard to form by pure chance and located in protein disordered regions) studied by Hagai et al. [87], tend to have a smaller size than sets $C \cap D$, of SLiMs conserved and located in protein disordered regions.

The intersection SLiM sets $C \cap R$ (conserved and rare) and $C \cap D \cap R$ (conserved, rare, located in disordered regions) are almost empty so they can be discarded as useful filtering criteria, data not shown.

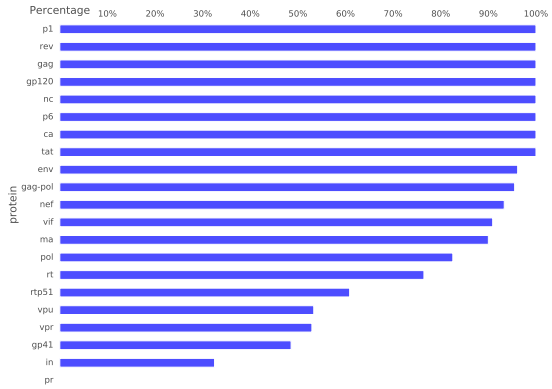


Figure 7: Percentage of conserved motifs that are located in disordered regions in HIV-1.

The proteins that have less conserved and disordered content are gp41, in and pr. For in (integrase) and pr (protease) disordered regions, see Figure 6.

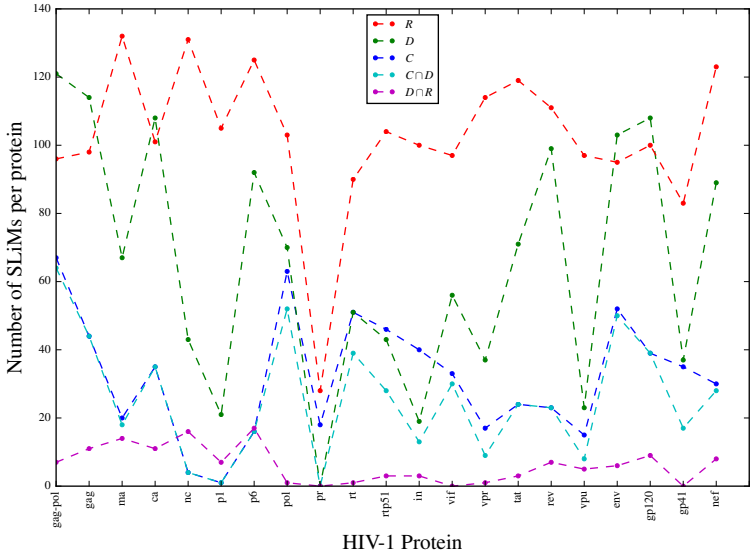


Figure 8: Number of SLiMs by set.

We plot the number of SLiMs (regular expressions) that were found in the HIV-1 proteins. The intersection SLiM sets $C \cap R$ (conserved and rare) and $C \cap D \cap R$ (conserved, rare, located in disordered regions) were discarded for being almost null in all entries.

7.4 PROTEIN-PROTEIN INTERACTIONS PREDICTED WITH THE SLIM SETS ARE ENRICHED IN EXPERIMENTALLY VALIDATED HIV-1-HUMAN PROTEIN-PROTEIN INTERACTIONS

We validate against two virus-host PPIs databases: NIAID [79] and LMPID [174]. The NIAID contains 15074 PPIs at the moment of writ-

ing while LMPID contains 2203 PPIs between several viruses and hosts, with 6 PPIs between HIV-1 and human.

The validation of the predicted PPIs with the NIAID database is not the best way to gauge the proportion of SLiM-based interactions. This database contains PPIs of all kinds, not only SLim-mediated ones. However, it is the most complete virus-host PPI dataset.

A better validation set, conceptually, is constructed with pairs deemed to interact through a SLiM with the LMPID database. Nevertheless, this dataset is too small. We perform the comparison with both databases, selecting the NIAID database to compare the sets prediction performance and check the statistical significance of the results.

Although we are suggesting a partial ranking of SLiM-based predicted PPIs, another addition would be to rank totally the interactions with a score representing the probability that the interaction takes place based on experimental data [22] or other techniques [150]. For the moment, a total ranking is difficult to achieve given the scarcity of data about SLiM-mediated PPIs [174, 18].

7.4.1 *In the NIAID database*

If we validate against the NIAID database of HIV-1-human PPIs we find the number of interactions reported in Table 6. The SLiM sets have a general tendency with respect to the number of PPIs correctly predicted, across all HIV-1 proteins, Figure 10. For this reason we propose to rank the PPIs predicted according to the set used to deduce them. The sets are ranked partially by the number of correctly predicted PPIs in the NIAID database.

The p-values for the overlap between the PPIs predicted with base on each SLiM set and the PPIs in the NIAID database are in Table 7.

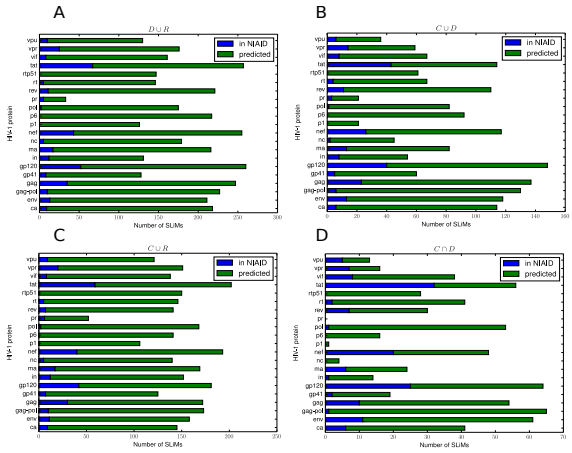


Figure 9: Number of hits per set.

The number of true positives, correctly inferred PPIs, as validated with the NIAID HIV-1 Human Interaction Database. The number of predicted interactions is represented with a green bar. The blue bar represents the number of interactions predicted and validated.

Table 6: Number of predicted interactions per SLiM set that have experimental support in NIAID database.

	C	D	R	$C \cup D$	$C \cup R$	$D \cup R$	$C \cup D \cup R$	$C \cap D$	$D \cap R$
ca	6	6	6	6	9	9	9	6	1
env	11	13	6	13	11	13	13	11	0
gag-pol	0	6	10	6	10	10	10	0	5
gag	10	23	27	23	30	35	35	10	10
gp41	4	3	7	5	7	8	8	2	0
gp120	25	40	37	40	42	52	52	25	0
in	8	1	12	8	12	12	12	1	0
ma	6	13	16	13	17	17	17	6	11
nc	0	2	5	2	5	5	5	0	2
nef	20	26	34	26	40	43	43	20	0
p1	0	0	0	0	0	0	0	0	0
p6	0	0	0	0	0	0	0	0	0
pol	1	1	2	1	2	2	2	1	0
pr	3	0	5	3	6	5	6	0	0
rev	7	11	7	11	7	11	11	7	2
rt	4	2	5	4	5	5	5	2	0
rtp51	0	0	0	0	0	0	0	0	0
tat	32	43	53	43	59	67	67	32	3
vif	8	8	6	8	8	8	8	8	0
vpr	7	14	20	14	20	25	25	7	2
vpu	5	6	9	6	9	10	10	5	4

Table 7: Overlap between predicted interactions and NIAID PPI database.

	C	D	R	$C \cup D$	$C \cup R$	$D \cup R$	$C \cup D \cup R$	$C \cap D$	$D \cap R$
ca	0.00507	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00507	0.01714
env	0.00004	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00004	0.00255
gag-pol	0.00000	0.00000	0.00001	0.00000	0.00000	0.00000	0.00000	0.00000	0.45495
gag	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.06194
gp41	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
gp120	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
in	0.01407	0.00149	0.00037	0.01273	0.00006	0.00019	0.00006	0.00180	0.75138
ma	0.03339	0.10909	0.00486	0.10360	0.00358	0.00242	0.00242	0.07022	0.27587
nc	0.64494	0.00133	0.00117	0.00133	0.00111	0.00091	0.00091	0.64494	0.01188
nef	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
p1	0.00000	0.72014	0.45345	0.72014	0.45345	0.43570	0.43570	0.00000	0.83020
p6	0.34157	0.07552	0.01970	0.07552	0.01653	0.00998	0.00998	0.34157	0.38820
pol	0.01830	0.01095	0.00697	0.00720	0.00164	0.00091	0.00080	0.02737	0.83859
pr	0.00000	0.00000	0.00009	0.00000	0.00000	0.00009	0.00000	0.00000	0.00000
rev	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02941
rt	0.06943	0.00454	0.00258	0.01358	0.00033	0.00034	0.00020	0.03287	0.00000
rtp51	0.75895	0.77358	0.61184	0.70277	0.56852	0.56784	0.55594	0.83653	0.99740
tat	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.35524
vif	0.00015	0.00002	0.00000	0.00002	0.00000	0.00000	0.00000	0.00016	0.00000
vpr	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.83217
vpu	0.00464	0.00277	0.00000	0.00092	0.00000	0.00000	0.00000	0.02055	0.03182

The p-value indicates the probability that the overlap between our sets of predicted PPIs and the PPIs with literature support in the NIAID database takes place under the null hypothesis, that our sets were formed by random sampling. Red values are not significant at a level of 0.05.

7.4.2 *In the LMPID database*

Taking as control the literature curated PPIs mediated by SLiMs between HIV-1 and human proteins we find that the motif sets $C, C \cap D, C \cap R, C \cap D \cap R$ capture half of them, while the sets $C \cup D, C \cup R, D \cup R, C \cup D \cup R, D \cup R$ allow to infer all of them [174].

The small number of human-HIV-1 interactions in this database, six, leaves open two possibilities: the number is really small, or the number is larger but few experiments have been performed to detect them. Indeed, to estimate the number of human-HIV-1 SLiM-

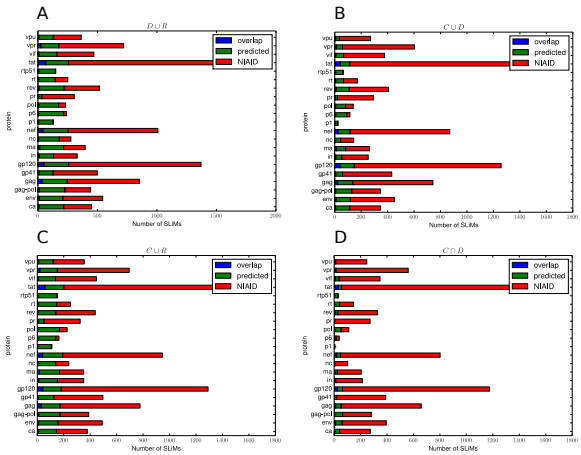


Figure 10: Fraction of hits per set.
The number of hits, correctly inferred PPIs, as validated with the NIAID HIV-1 Human Interaction Database. The number of predicted interactions is represented with the green bar. The blue bar represents the number of interactions predicted and validated. The red bar represents the number of interactions in the NIAID database.

mediated PPIs more work is needed, perhaps an approach based on combining expert opinions [189].

7.4.3 *Interaction listings*

The interactions inferred for each protein with each SLiM set are in files named in the following manner: hivProtein_interactions_SLiMset.csv. Where hivProtein is one of ca, env, gag-pol, gag, gp41, etc. And SLiMset is one of suffixes listed in Table 8. Each entry in the csv files is a Uniprot id for the human protein interacting with the HIV-1 protein that names the file. The interactions that we inferred and are in the NIAID database are counted in Table 6 and are in supplementary file `validatedinteractions.zip`. The interactions inferred not validated, the candidates for novel interactions, are in the supplementary file `interactions.zip` and counted in Table 9.

Table 8: Suffixes used in the filenames of the interaction sets

SLiM set	Suffix
Conserved [68]	C
Disordered [87]	D
Rare [87]	R
Conserved and disordered	iCD
Conserved and rare	iCR
Disodered and rare	iDR
Conserved and disordered and rare	iCDR
Conserved or disordered	uCD
Conserved or rare	uCR
Disodered or rare	uDR
Conserved or disordered or rare	uCDR

7.4.4 *Sensitivity and specificity*

Although there is no gold-standard dataset for VHPPIs we use the NIAID database to estimate the sensitivity of the SLiM-based predictions. We iterate through all possible interactions between human and HIV-1 proteins to compute the true positives, true negatives, false positives and false negatives. Tables 10 and 11 report the sensitivity and specificity. The values are discriminated per HIV-1 protein and SLiM set used to infer the intearcions.

7.5 PPIS CORRECTLY PREDICTED SERVE AS A RANKING OF FILTERING METHODS

Taking as validation set the bigger database of interactions, the NIAID, the ranking of sets per the number of PPIs deduced that match a database record is $D \cup R, C \cup R, C \cup D, C \cap D, D \cap R$, see figure 11

The ranking of these sets allow to present the PPIs predicted to researchers in a partial order by two blocks, each one with PPIs deduced with the sets $D \cup R, C \cup R$.

Table 9: Interactions predicted and not validated in the NIAID database

protein	C	D	R	$C \cup D$	$C \cup R$	$D \cup R$	$C \cup D \cup R$	$C \cap D$	$C \cap R$	$D \cap R$
ca	3045	6483	5743	6483	7319	9008	9008	3045	0	1027
env	4382	5999	5648	6033	7730	9028	9061	4348	0	598
gag-pol	5255	7179	5139	7326	8192	9339	9405	5108	0	199
gag	4237	6794	5386	6794	7614	9139	9139	4237	0	1030
gp41	3735	4067	5208	4851	6617	6908	7497	2751	0	0
gp120	3819	6058	5419	6058	7100	8832	8832	3819	0	678
in	3828	1613	6323	3915	7606	6605	7645	1476	0	62
ma	2922	5134	7007	5175	7955	8752	8760	2692	24	2654
nc	194	4462	7533	4462	7578	8146	8146	194	0	2971
nef	3375	5793	6875	5826	8053	9672	9691	3341	0	1017
p1	0	2771	6967	2771	6967	7497	7497	0	0	1417
p6	2008	5492	6910	5492	7552	9365	9365	2008	0	1465
pol	4854	5354	5917	5703	8377	8840	9025	4498	0	140
pr	2744	0	2937	2744	4069	2937	4069	0	0	0
rev	3282	6340	6143	6340	7052	9044	9044	3282	0	935
rt	4075	4288	5638	4985	7762	7753	8126	3366	0	50
rtp51	4115	4003	6482	5220	8269	8274	8772	2825	0	113
tat	2753	4928	6545	4928	7259	8932	8932	2753	0	159
vif	3284	4026	5583	4075	6766	7414	7442	3215	0	0
vpr	2312	3381	6922	3920	7696	7989	8491	1723	0	96
vpu	2420	2916	5769	3353	6375	6426	6690	1895	0	1439

The number of interactions predicted is high in many cases. The reason is the number of proteins and isoforms that contain a domain that is deemed to interact with a SLiM.

Table 10: Sensitivity percentage for SLiM sets prediction over HIV-1 proteins

protein	C	D	R	$C \cup D$	$C \cup R$	$D \cup R$	$C \cup D \cup R$	$C \cap D$	$D \cap R$
ca	1.99	3.90	3.82	3.90	4.54	5.16	5.16	1.99	0.75
env	2.12	2.97	3.13	2.97	3.79	4.42	4.42	2.12	0.19
gag-pol	3.67	4.05	3.07	4.45	4.82	5.03	5.14	3.27	0.09
gag	2.21	3.16	2.26	3.16	3.31	3.83	3.83	2.21	0.50
gp41	2.30	2.58	3.03	2.97	3.70	3.91	4.13	1.81	0.00
gp120	1.17	1.85	1.84	1.85	2.12	2.60	2.60	1.17	0.23
in	2.56	1.72	3.86	2.65	4.45	4.28	4.48	1.63	0.08
ma	1.74	2.78	3.77	2.83	4.21	4.57	4.57	1.35	1.36
nc	0.18	3.00	4.24	3.00	4.24	4.60	4.60	0.18	2.33
nef	1.14	2.31	2.49	2.31	2.62	3.28	3.28	1.14	0.41
p1	0.00	2.04	4.55	2.04	4.55	4.82	4.82	0.00	1.53
p6	1.52	3.36	4.84	3.36	5.07	5.80	5.80	1.52	1.17
pol	3.59	3.43	4.03	3.88	5.57	5.61	5.76	3.14	0.01
pr	1.48	0.00	2.02	1.48	2.57	2.02	2.57	0.00	0.00
rev	2.08	3.79	3.99	3.79	4.44	5.27	5.27	2.08	0.21
rt	2.75	2.60	3.46	3.01	4.54	4.51	4.69	2.34	0.00
rtp51	3.09	2.64	4.41	3.60	5.41	5.38	5.63	2.11	0.09
tat	0.98	1.48	2.17	1.48	2.39	2.73	2.73	0.98	0.05
vif	2.01	2.42	3.22	2.42	3.90	4.19	4.19	1.94	0.00
vpr	0.81	1.65	3.11	1.86	3.36	3.59	3.81	0.59	0.06
vpu	1.34	1.72	2.75	1.95	2.89	3.28	3.34	0.93	1.00

Table 11: Specificity percentage for SLiM sets prediction over HIV-1 proteins

protein	C	D	R	$C \cup D$	$C \cup R$	$D \cup R$	$C \cup D \cup R$	$C \cap D$	$D \cap R$
ca	98.87	97.62	97.4	97.62	96.91	96.41	96.41	98.87	99.61
env	98.40	97.74	97.41	97.72	96.84	96.41	96.40	98.42	99.71
gag-pol	97.98	97.38	97.6	97.28	96.59	96.36	96.31	98.08	99.94
gag	98.41	97.57	97.56	97.57	96.86	96.43	96.43	98.41	99.61
gp41	98.58	98.34	97.85	98.09	97.47	97.32	97.15	98.91	100.00
gp120	98.65	97.77	97.62	97.77	97.19	96.6	96.60	98.65	99.69
in	98.53	99.3	97.48	98.51	97.09	97.36	97.09	99.33	99.97
ma	98.84	98.08	97.19	98.06	96.91	96.74	96.73	98.96	98.79
nc	99.91	98.25	97.09	98.25	97.07	96.91	96.91	99.91	98.79
nef	98.72	97.87	97.06	97.87	96.74	96.25	96.25	98.72	99.59
p1	100.00	98.9	97.18	98.9	97.18	97.00	97.00	100.00	99.40
p6	99.24	97.96	96.97	97.96	96.85	96.33	96.33	99.24	99.37
pol	98.13	97.94	97.35	97.76	96.59	96.44	96.36	98.31	99.90
pr	98.93	100.00	98.86	98.93	98.48	98.86	98.48	100.00	100.00
rev	98.60	97.65	97.23	97.65	96.9	96.44	96.44	98.60	99.58
rt	98.46	98.41	97.47	98.06	96.82	96.89	96.69	98.80	99.99
rtp51	98.38	98.48	97.16	97.95	96.59	96.63	96.42	98.96	99.98
tat	98.87	98.18	97.28	98.18	97.06	96.55	96.55	98.87	99.95
vif	98.74	98.52	97.52	98.50	97.20	96.99	96.99	98.76	100.00
vpr	99.09	98.71	97.04	98.53	96.8	96.75	96.58	99.27	99.95
vpu	99.03	98.81	97.5	98.66	97.34	97.32	97.25	99.23	99.37

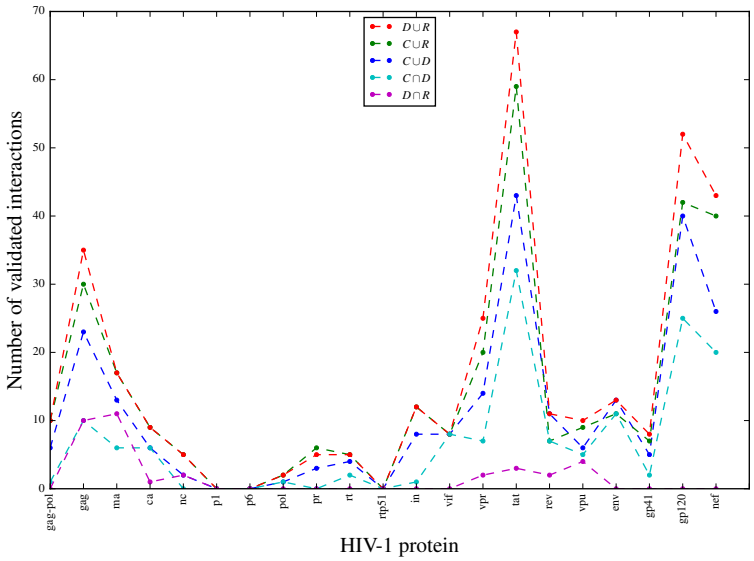


Figure 11: Comparison of sets by number of hits.

The number of hits, correctly inferred PPIs, as validated with the NIAID HIV-1 Human Interaction Database. Each set is represented by a line with a different color in order to find a tendency in the majority of the proteins. The number of hits allows to order the sets in this way: $D \cup R$, $C \cup R$,$C \cup D$,$C \cap D$ and $D \cap R$.

8

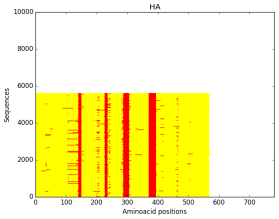
ANALYSIS OF THE SLIM-MEDIATED VIRAL SUBVERSION MECHANISMS OF THE HOST PROTEIN-SYNTHESIS MACHINERY

8.1 DISORDER IN VIRAL PROTEINS

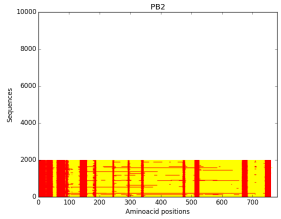
In all the figures the yellow line in the plots represents a protein sequence. The red segments denote disordered regions as deduced with IUPred with the sliding window addition explained in Chapter 5, section 5.2.1. We show a subset of all the viral proteins analyzed.

8.1.1 *Influenza AH1N1*

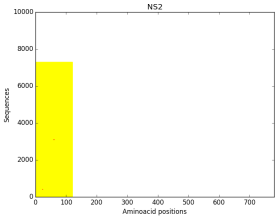
Disorder for influenza virus AH1N1 proteins HA (hemagglutinin), PB2, NS2 and M1 is presented in Figure 12.



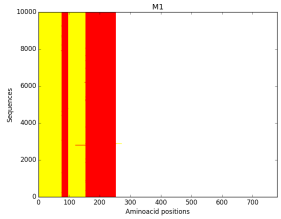
(a) HA (hemagglutinin) is a glycoprotein used in attachment and entry of the virion.



(b) PB2 is a component of the viral polymerase.



(c) NS2 is a protein involved in nuclear export of viral complexes. It is the most ordered protein for AH1N1.

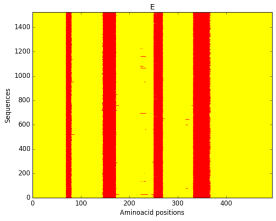


(d) M1, a matrix protein.

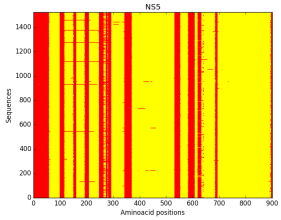
Figure 12: Disorder in Influenza A H1N1 proteins.

8.1.2 *Dengue-1*

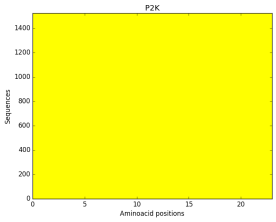
Disorder for proteins E, NS5, P2K and C is presented in Figure 13.



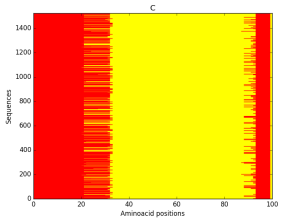
(a) E stands for envelope protein and is involved in attachment and entry of the virion.



(b) NS5 is the RNA-directed RNA polymerase.



(c) P2K is a signal peptide involved in interferon antagonism.

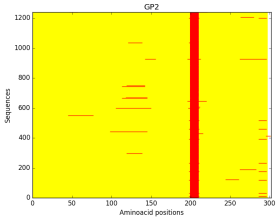


(d) C, the capsid protein.

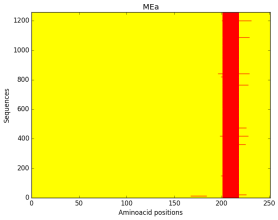
Figure 13: Disorder in Dengue-1 proteins.

8.1.3 *Ebola*

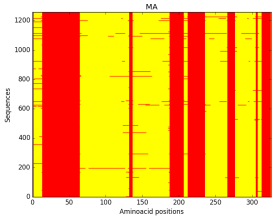
Disorder for proteins Ebola proteins GP2, MeA, MA and NP is presented in Figure 14.



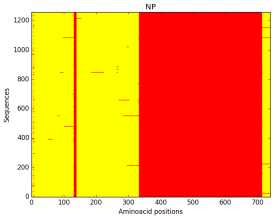
(a) GP2 is a glycoprotein involved in attachment and entry of the virion.



(b) MEa is the membrane associated protein. Together with GP2, these are the most ordered proteins in Ebola.



(c) The matrix protein

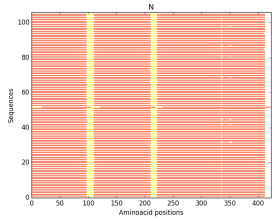


(d) The nucleoprotein.

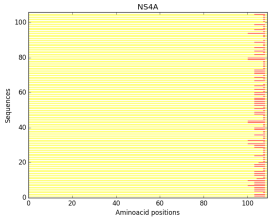
Figure 14: Disorder in Ebola(Zaire) proteins.

8.1.4 *MERS*

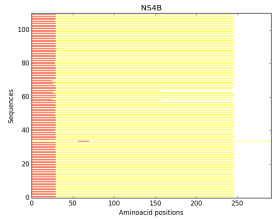
Disorder for MERS viral proteins N, NS4A, NS4B, and NS5 is presented in Figure 15.



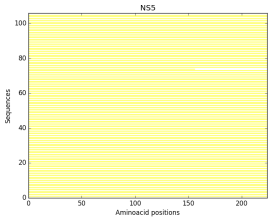
(a) N is the nucleocapsid protein.



(b) NS4A inhibits of host innate responses.



(c) NS4B also inhibits innate responses.

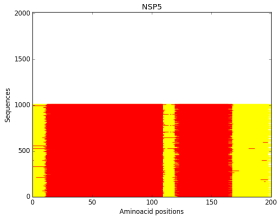


(d) NS5 might interfere with the interferon system.

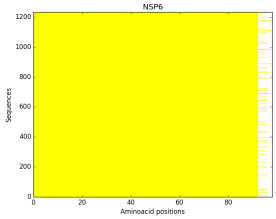
Figure 15: Disorder in MERS coronavirus proteins.

8.1.5 *Rotavirus*

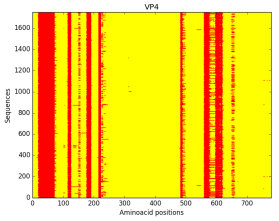
Disorder for proteins NSP5, NSP6, VP4 and VP7 is presented in Figure 16.



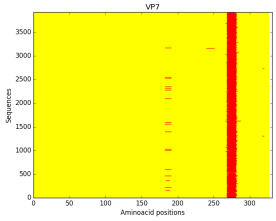
(a) NSP5 contributes to genome replication.



(b) NSP6 contributes to virion packaging.



(c) VP4 contributes to virion attachment to the cell membrane.

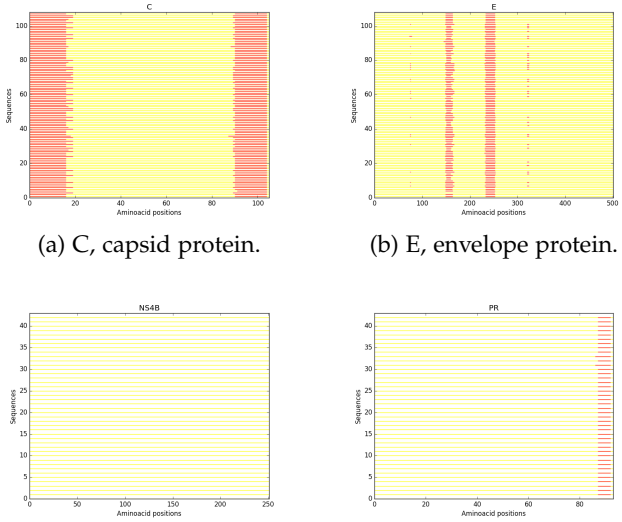


(d) VP7 also contributes to attachment.

Figure 16: Disorder in Rotavirus proteins.

8.1.6 *West Nile*

Disorder for West Nile virus proteins C, E, NS4B and NS5 is presented in Figure 17.



(a) C, capsid protein.

(b) E, envelope protein.

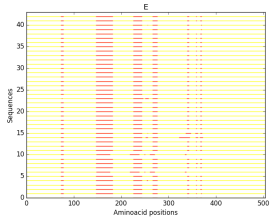
(c) NS4B antagonizes the interferon system.

(d) NS5 is the RNA-directed RNA polymerase.

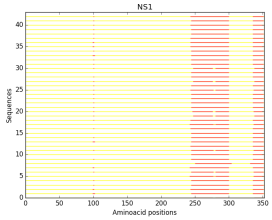
Figure 17: Disorder in West Nile virus proteins.

8.1.7 *Zika*

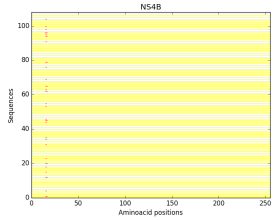
Disorder for Zika virus proteins E, NS1, NS4B and PR is presented in Figure 18.



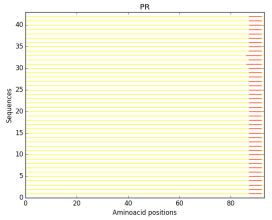
(a) E, envelope protein.



(b) NS1 is associated with virus replication and regulation of the innate immune response.



(c) NS4B antagonizes the interferon system.



(d) PR is a peptyde with unknown functions.

Figure 18: Disorder in Zika virus proteins.

Table 12: Number of interactions with viral proteins

Human HPSM protein	Interactions with viral proteins
EIF4A1	30
EIF4A2	30
EIF4A3	30
EIF3B	44
EIF3G	44
PABPC5	44
PABPC1	50
PABPC3	50
PABPC4	50
EIF4E	55
EIF4E1B	55
EIF4E2	55
EIF4E3	55
EIF3I	78

8.2 TARGETED PROTEINS

There are only two kinds of human proteins in the HPSM targeted by the selected viruses: 1) eukaryotic Initiation Factors (EIF*), 2) polyadenilate-binding proteins (PABPC*). No cytoplasmic ribosomal proteins or components of the ribosomal units are predicted to interact with the viral proteins. The number of interactions with viral proteins for the targeted proteins is reported in Table 12

Targeted proteins EIF3B, EIF3G and EIF3I belong to the module A of the EIF3 complex involved in the recruitment of the 43S ribosomal complex at the translation initiation phase.

Proteins EIF4A1, EIF4A2, EIF4A3, EIF4E, EIF4E1B, EIF4E2 and EIF4E3 are part of the EIF4 complex that binds to capped mRNAs in the translation initiation phase.

Finally, proteins PAPBPC1, PAPBPC3, PAPBPC4 and PAPBPC5 bind to the tail (end) of mRNAs recognizing poly(A) regions. This helps to mRNA circularization.

8.3 VIRUS-HOST PROTEIN-PROTEIN INTERACTIONS

The SLiM-mediated VHPPI prediction method finds 670 interactions between HPSM proteins and viral proteins with the selected viruses. The resulting network is presented in Figure 19.

We present two degree distributions for the network, one for the human proteins with respect to the number of interactions with viral

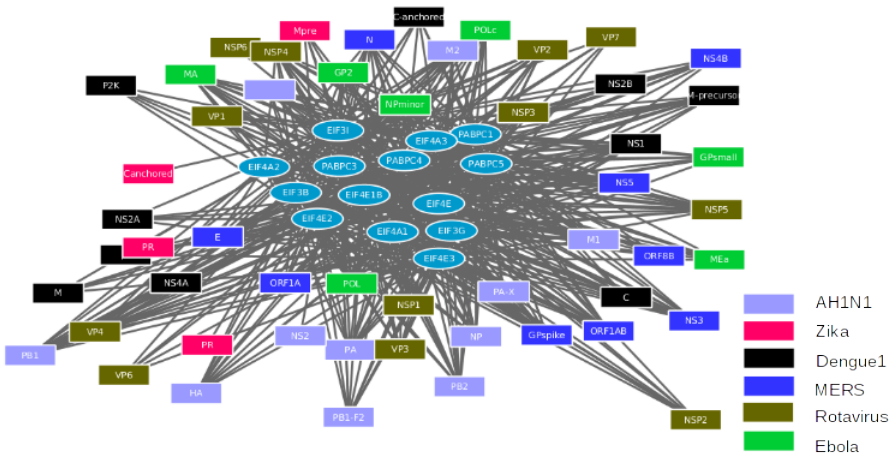


Figure 19: Network for HPSM and viral proteins
Human protein-synthesis proteins are represented as ellipses and viral proteins as boxes. Boxes are colored differently for each virus.

Table 13: Degree distribution for human proteins

Human protein degree	Number of proteins
30	3
44	3
50	3
55	4
78	1

proteins in Table 13, and other for viral proteins with respect to the number of interactions with human proteins in Table 14. For human proteins there is a clear hub, the protein EIF3I, predicted to interact with 78 viral proteins through SLiMs, but the other proteins have a large degree,- Table 13. On the other hand, there are 27 viral proteins that have 14 interactions with human proteins.

We classify the viral proteins in two groups: 1) the ones that have only one interaction with human proteins, potentially disrupting the protein-synthesis process and 2) the ones that have two or more interactions with human proteins, potentially bridging unexpected interactions between human protein-synthesis proteins. These first group of potentially disrupting proteins is presented in Table 15 and the viral hubs are presented in Table 16.

Other viral proteins with more than one interaction with human proteins are listed in Table 17.

Table 14: Degree distribution for viral proteins

Viral Degree	Number of proteins
1	16
4	1
5	11
7	5
8	5
10	1
11	11
14	27

Table 15: Viral proteins with one interaction (potentially disrupting)

Virus	Viral protein
Zika	PR
MERS	NS5
Rotavirus	NSP6
West Nile	C
Dengue1	NS4B
Dengue1	NS4A
West Nile	C-anchored
Dengue1	M
West Nile	M
Dengue1	NS2A
MERS	E
West Nile	NS4A
West Nile	NS4B
Zika	C
AH1N1	NS2
Zika	Canchored

Table 16: Viral hub proteins (potentially bridging)

Virus	Protein
AH1N1	NP
AH1N1	M1
AH1N1	NS1
AH1N1	PA
AH1N1	PB1
AH1N1	PB2
Dengue1	NS3
Dengue1	NS5
Ebola	GPspike
Ebola	MA
Ebola	NP
Ebola	NPminor
Ebola	POL
Ebola	POLc
MERS	N
MERS	ORF1AB
Rotavirus	NSP4
Rotavirus	NSP5
Rotavirus	VP2
Rotavirus	VP4
West Nile	E
West Nile	NS1
West Nile	NS3
West Nile	NS5
Zika	NS1
Zika	NS3

Table 17: Potentially bridging viral proteins

Virus	Protein	Interactions
Dengue1	NS2B	4
MERS	NS3	5
Zika	NS4B	5
Zika	Mpre	5
Rotavirus	VP6	5
Rotavirus	VP7	5
Dengue1	pr	5
WestNile	M-precursor	5
WestNile	NS2A	5
MERS	GPspike	5
Dengue1	P2K	5
Zika	M	5
Dengue1	NS1	7
Dengue1	M-precursor	7
MERS	ORF8B	7
Dengue1	C-anchored	7
Dengue1	C	7
Rotavirus	VP1	8
Rotavirus	NSP2	8
MERS	NS4B	8
Ebola	GPsmall	8
Ebola	MEa	8
Rotavirus	NSP1	10
AH1N1	HA	11
AH1N1	PB1-F2	11
Ebola	GP2	11
AH1N1	PA-X	11
Rotavirus	VP3	11
Dengue1	E	11
Zika	E	11
AH1N1	M2	11
MERS	ORF1A	11
Zika	NS5	11
Rotavirus	NSP3	11

Part IV

CONCLUSIONS AND FUTURE WORK

9 CONCLUSIONS AND FUTURE WORK

9.1 CONCLUSIONS

We develop a bioinformatic method to predict virus-host SLiM-mediated PPIs and rank them. It is applicable to any eukaryotic virus and host with available protein sequences. The requirements for the method are: 1) the availability of a reference genome for the virus, 2) the availability of viral sequences, 3) protein-domain associations for the host organism, 4) the ELM-database of SLiMs.

Using data for the most studied virus, HIV-1, we find a partial ordering of the PPIs obtained based on the set used to infer the interactions. The order consists in two blocks: $D \cup R, C \cup R$. This order is descending in the expected probability of inferring real interactions. We expect that the method gives interesting candidate interactions with other eukaryotic viruses and hosts.

Most of the HIV-1 conserved motifs are located in disordered regions, suggesting that protein structural flexibility could be an important factor to accommodate SLiMs to mimic host proteins.

Although there are machine learning methods for predicting host-pathogen PPIs [11, 63], the descriptors they use are primarily based on domain-domain interactions, not on motif-domain interactions. We consider our approach as a different way to obtain candidate interactions and rank them.

The call for using high-throughput methods to detect SLiM-mediated PPIs illustrates the benefits of a bioinformatic method that predicts SLiM-mediated PPIs and might guide experimental design [18].

9.2 FUTURE WORK

The recent study of fuzziness and SLiM flanking regions opens a window to understand more the nature of SLiM-mediated PPIs [62]. Advances in this direction might result in better SLiM filtering methods.

An additional filter to consider is if the protein binding region in the host protein interacting with a SLiM is disordered, as deemed by a predictor like ANCHOR [132].

Another possible addition is a filtering method based on structural properties of the SLiM like being exposed at the protein surface like it is done in [176]. Indeed, there is a previous work that proposes to extend the notion of SLiM to include a structural component [173].

We performed the prediction of SLiM-mediated host-virus PPIs between the human HPSM and some selected viruses. However, the

methods proposed can be extended to other subsystems like the cell entry [182], interferon [141], complement [16], apoptosis proteins [89], the nucleus [35] and the cytoskeleton [202] to investigate viral infection mechanisms at different stages of the infectious cycle.

Part V

APPENDIX

A

COMPUTATIONAL ASPECTS

A.1 SOFTWARE ENGINEERING

Being a small development the scripts developed during the research do not follow a particular software engineering methodology. However, we have been careful to write command-line scripts with a simple input-output transformation and a straightforward documentation.

The tasks are divided in processing stages that take inputs like protein sequences, DNA sequences, protein alignments and IUPred outputs stored in a particular format –Fasta, IUpred output in text files – and compute an output in some specific format – CSV, Graphml.

All the scripts are written with the same structure: a header explaining the purpose of the script (input-output transformation), then a series of functions and classes, followed by a __main__ python section containing a small test case of the script, with hard-coded inputs, documenting its mode of use. See an example in section A.3.

A.2 PROGRAMMING ENVIRONMENT

All the scripts are written in Python (version 3) programming language. The libraries used are pandas (for data processing), biopython (for bioinformatics formats and algorithms), networkx (for network generation), matplotlib (for graph generation) and scipy (for statistical computations).

A.3 EXAMPLE: GAG-POL TRANSLATION

The NIAID does not allow to download translated gag-pol sequences and the translation is not immediate because of the frame-shift that permits the overlapping of gag and pol sequences in different reading frames. That led us to develop a simple translation in the file translateGagPol.py, reproduced below:

Listing 1: Translation of HIV-1 Gag-Pol ADN to protein

```
1 #!/usr/bin/env python3
  # Author: Andres Becerra sandoval <andres.becerra at gmail.com>
  # Tested in python-3.4.3

  # translateGagPol.py translates gag-pol from DNA to protein
6 # input: multiple DNA sequences in a fasta file
  # output: multiple protein sequences in a fasta file
```

```
import re
import sys
from Bio import SeqIO
from Bio.Seq import Seq
from Bio.Alphabet import IUPAC

def translateGagPol(proteinFile, result):
    fout = open(result, 'w')
    for record in SeqIO.parse(proteinFile, 'fasta'):
        content = str(record.seq)
        regex = re.compile("TTTTTTA")
        r = regex.search(content)

        if r != None:
            ini = r.start()
            end = r.end()

            a = record.seq[:end].translate(to_stop=False)
            b = record.seq[end-1:].translate(to_stop=False)
            gag_pol = Seq("", IUPAC.protein)
            gag_pol = a+b
            fout.write('>'+record.id+'\n')
            lines = [str(gag_pol[i:i+70]) for i in range(0,
                                                len(gag_pol),70)]
            for line in lines:
                fout.write(line+'\n')
    fout.close()

if __name__ == '__main__':
    if len(sys.argv) != 3:
        print('Usage: translateGagPol.py proteinFile.fasta'
              ' output.fasta')
        print('Example:')
        path = '/home/abecerra/doctorado/data/'
        proteinFile = path + 'hiv-sequences/HIV1_less2015_gagpol_DNA.
            fasta'
        result = "/tmp/out.fasta"
        print('translateGagPol.py ' + proteinFile + ' ' + result)
    else:
        proteinFile = sys.argv[1]
        result = sys.argv[2]

    translateGagPol(proteinFile, result)
```

The header takes place between lines 1-7, library imports are in lines 8-13, the function that computes the translation is in lines 15-35 and the main section, that makes the script a command line executable, is in lines 37-50. Notice that the main section documents how to use the script from the command line and from a different python module. All the scripts developed follow the same structure.

BIBLIOGRAPHY

[1] M. Ahmad, K. Pyaram, J. Mullick, and A. Sahu. Viral complement regulators: the expert mimicking swindlers. *Indian J. Biochem. Biophys.*, 44(5):331–343, Oct 2007. (Cited on page 1.)

[2] D. Ako-Adjei, W. Fu, C. Wallin, K. S. Katz, G. Song, D. Darji, J. R. Brister, R. G. Ptak, and K. D. Pruitt. HIV-1, human interaction database: current status and new features. *Nucleic Acids Res.*, 43(Database issue):D566–570, Jan 2015. (Cited on page 17.)

[3] A. Alcami. Viral mimicry of cytokines, chemokines and their receptors. *Nat. Rev. Immunol.*, 3(1):36–50, Jan 2003. (Cited on pages 1 and 23.)

[4] T. Altman, M. Travers, A. Kothari, R. Caspi, and P. D. Karp. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, 14:112, 2013. (Cited on page 28.)

[5] L. A. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. U.S.A.*, 97(21):11149–11152, Oct 2000. (Cited on page 30.)

[6] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, 36(Database issue):D419–425, Jan 2008. (Cited on page 20.)

[7] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000. (Cited on pages 20 and 27.)

[8] G. D. Bader, D. Betel, and C. W. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 31(1):248–250, Jan 2003. (Cited on pages 25 and 26.)

[9] S. M. Bailer and J. Haas. Connecting viral with cellular interactomes. *Curr. Opin. Microbiol.*, 12(4):453–459, Aug 2009. (Cited on page 24.)

[10] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL http://www.sciencemag.org/content/286/5439/509.abstract. (Cited on pages 27 and 31.)

[11] R. K. Barman, S. Saha, and S. Das. Prediction of interactions between viral and host proteins using supervised machine learning methods. *PLoS ONE*, 9(11):e112034, 2014. (Cited on pages 3 and 71.)

[12] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. D. Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L. G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. MacDougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, B. E. Suzek, C. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, M. S. Yerramalla, and J. Zhang. UniProt: a hub for protein information. *Nucleic Acids Res.*, 43(Database issue):D204–212, Jan 2015. (Cited on pages 7 and 39.)

[13] A. Ben-Hur and W. S. Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1:S2, 2006. (Cited on page 3.)

[14] R. B. Berlow, H. J. Dyson, and P. E. Wright. Functional advantages of dynamic protein disorder. *FEBS Lett.*, 589(19 Pt A): 2433–2440, Sep 2015. (Cited on page 21.)

[15] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, 58(Pt 6 No 1):899–907, Jun 2002. (Cited on pages 21 and 25.)

[16] J. Bernet, J. Mullick, A. K. Singh, and A. Sahu. Viral mimicry of the complement system. *J. Biosci.*, 28(3):249–264, Apr 2003. (Cited on page 72.)

[17] A. Bet, E. A. Maze, A. Bansal, S. Sterrett, A. Gross, S. Graff-Dubois, A. Samri, A. Guihot, C. Katlama, I. Theodorou, J. M. Mesnard, A. Moris, P. A. Goepfert, and S. Cardinaud. The HIV-1 antisense protein (ASP) induces CD8 T cell responses during chronic infection. *Retrovirology*, 12:15, 2015. (Cited on page 14.)

[18] C. Blikstad and Y. Ivarsson. High-throughput methods for identification of protein-protein interactions involving short linear motifs. *Cell Commun. Signal*, 13:38, 2015. (Cited on pages 25, 48, and 71.)

[19] J. S. Bonifacino, E. C. Dell'Angelica, and T. A. Springer. Immunoprecipitation. *Curr Protoc Mol Biol*, Chapter 10:Unit 10.16, May 2001. (Cited on page 24.)

[20] Mehdi Bouraí, Marianne Lucas-Hourani, Hans Henrik Gad, Christian Drosten, Yves Jacob, Lionel Tafforeau, Patricia Cassonnet, Louis M. Jones, Delphine Judith, Thérèse Couderc, Marc Lecuit, Patrice André, Beate Mareike Kümmerer, Vincent Lotteau, Philippe Desprès, Frédéric Tangy, and Pierre-Olivier Vidalain. Mapping of chikungunya virus interactions with host proteins identified nsp2 as a highly connected viral component. *Journal of Virology*, 86(6):3121–3134, 2012. doi: 10.1128/JVI.06390-11. URL http://jvi.asm.org/content/86/6/3121.abstract. (Cited on page 31.)

[21] P. Braun. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. *Proteomics*, 12(10):1499–1518, May 2012. (Cited on page 24.)

[22] P. Braun, M. Tasan, M. Dreze, M. Barrios-Rodiles, I. Lemmens, H. Yu, J. M. Sahalie, R. R. Murray, L. Roncari, A. S. de Smet, K. Venkatesan, J. F. Rual, J. Vandenhaute, M. E. Cusick, T. Pawson, D. E. Hill, J. Tavernier, J. L. Wrana, F. P. Roth, and M. Vidal. An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods*, 6(1):91–97, Jan 2009. (Cited on pages 24 and 48.)

[23] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova. NCBI viral genomes resource. *Nucleic Acids Res.*, 43(Database issue):D571–577, Jan 2015. (Cited on pages 17 and 39.)

[24] N. J. Buchkovich, Y. Yu, C. A. Zampieri, and J. C. Alwine. The TORrid affairs of viruses: effects of mammalian DNA viruses on the PI3K-Akt-mTOR signalling pathway. *Nat. Rev. Microbiol.*, 6(4):266–275, Apr 2008. (Cited on page 39.)

[25] Z. Burda, A. Krzywicki, O. C. Martin, and M. Zagorski. Motifs emerge from function in model gene regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.*, 108(42):17263–17268, Oct 2011. (Cited on page 27.)

[26] M. A. Calderwood, K. Venkatesan, L. Xing, M. R. Chase, A. Vazquez, A. M. Holthaus, A. E. Ewence, N. Li, T. Hirozane-Kishikawa, D. E. Hill, M. Vidal, E. Kieff, and E. Johannsen. Epstein-Barr virus and virus human protein interaction maps. *Proc. Natl. Acad. Sci. U.S.A.*, 104(18):7606–7611, May 2007. (Cited on pages 29 and 31.)

[27] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, 40 (Database issue):D742–753, Jan 2012. (Cited on page 28.)

[28] 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):D669–673, Jan 2009. (Cited on pages 17, 25, and 26.)

[29] A. Chatr-Aryamontri, B. J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, T. Reguly, A. Breitkreutz, A. Sellam, D. Chen, C. Chang, J. Rust, M. Livstone, R. Oughtred, K. Dolinski, and M. Tyers. The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, 41(Database issue):D816–823, Jan 2013. (Cited on pages 25 and 26.)

[30] D. Chen, Y. Fong, and Q. Zhou. Specific interaction of Tat with the human but not rodent P-TEFb complex mediates the species-specific Tat activation of HIV-1 transcription. *Proc. Natl. Acad. Sci. U.S.A.*, 96(6):2728–2733, Mar 1999. (Cited on page 16.)

[31] K. C. Chen, T. Y. Wang, and C. H. Chan. Associations between HIV and human pathways revealed by protein-protein interactions and correlated gene expression profiles. *PLoS ONE*, 7(3): e34240, 2012. (Cited on page 17.)

[32] T. W. Chen, R. R. Gan, T. H. Wu, W. C. Lin, and P. Tang. VIP DB–a viral protein domain usage and distribution database. *Genomics*, 100(3):149–156, Sep 2012. (Cited on pages 17, 25, and 26.)

[33] U. Christen and M. G. von Herrath. Infections and autoimmunity–good or bad? *J. Immunol.*, 174(12):7481–7486, Jun 2005. (Cited on page 2.)

[34] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004. doi: 10.1103/PhysRevE.70.066111. URL http://link.aps.org/doi/10.1103/PhysRevE.70.066111. (Cited on page 28.)

[35] S. Cohen, S. Au, and N. Pante. How viruses access the nucleus. *Biochim. Biophys. Acta*, 1813(9):1634–1645, Sep 2011. (Cited on pages 16 and 72.)

[36] M. O. Collins and J. S. Choudhary. Mapping multiprotein complexes by affinity purification and mass spectrometry. *Curr. Opin. Biotechnol.*, 19(4):324–330, Aug 2008. (Cited on page 24.)

[37] M. J. Cowley, M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi, and J. Wu. PINA v2.0: mining interactome modules. *Nucleic Acids Res.*, 40(Database issue):D862–865, Jan 2012. (Cited on pages 25 and 26.)

[38] J. M. Cuevas, R. Geller, R. Garijo, J. Lopez-Aldeguer, and R. Sanjuan. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biol.*, 13(9):e1002251, Sep 2015. (Cited on page 4.)

[39] G. Cui, C. Fang, and K. Han. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*, 13 Suppl 7:S5, 2012. (Cited on page 3.)

[40] N. E. Davey, G. Trave, and T. J. Gibson. How viruses hijack cell regulation. *Trends Biochem. Sci.*, 36(3):159–169, Mar 2011. (Cited on pages 16 and 17.)

[41] N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson. Attributes of short linear motifs. *Mol Biosyst*, 8(1):268–281, Jan 2012. (Cited on pages 20 and 24.)

[42] N. E. Davey, M. S. Cyert, and A. M. Moses. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun. Signal*, 13(1):43, 2015. (Cited on page 23.)

[43] S. Davila, F. E. Froeling, A. Tan, C. Bonnard, G. J. Boland, H. Snippe, M. L. Hibberd, and M. Seielstad. New genetic associations detected in a host response study to hepatitis B vaccine. *Genes Immun.*, 11(3):232–238, Apr 2010. (Cited on page 19.)

[44] B. de Chassey, V. Navratil, L. Tafforeau, M. S. Hiet, A. Aublin-Gex, S. Agaugue, G. Meiffren, F. Pradezynski, B. F. Faria, T. Chantier, M. Le Breton, J. Pellet, N. Davoust, P. E. Mangeot, A. Chaboud, F. Penin, Y. Jacob, P. O. Vidalain, M. Vidal, P. Andre, C. Rabourdin-Combe, and V. Lotteau. Hepatitis C virus infection protein network. *Mol. Syst. Biol.*, 4:230, 2008. (Cited on pages 29 and 31.)

[45] B. de Chassey, L. Meyniel-Schicklin, A. Aublin-Gex, P. Andre, and V. Lotteau. New horizons for antiviral drug discovery from virus-host protein interaction networks. *Curr Opin Virol*, 2(5): 606–613, Oct 2012. (Cited on page 2.)

[46] B. de Chassey, A. Aublin-Gex, A. Ruggieri, L. Meyniel-Schicklin, F. Pradezynski, N. Davoust, T. Chantier, L. Tafforeau, P. E. Mangeot, C. Ciancia, L. Perrin-Cocon, R. Bartenschlager, P. Andre, and V. Lotteau. The interactomes of influenza virus NS1 and NS2 proteins identify new host factors and provide insights for ADAR1 playing a supportive role in virus replication. *PLoS Pathog.*, 9(7):e1003440, Jul 2013. (Cited on page 31.)

[47] J. De Las Rivas and C. Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.*, 6(6):e1000807, Jun 2010. (Cited on page 24.)

[48] E. Decroly, M. Vandenbranden, J. M. Ruysschaert, J. Cogniaux, G. S. Jacob, S. C. Howard, G. Marshall, A. Kompelli, A. Basak, and F. Jean. The convertases furin and PC1 can both cleave the human immunodeficiency virus (HIV)-1 envelope glycoprotein gp160 into gp120 (HIV-1 SU) and gp41 (HIV-I TM). *J. Biol. Chem.*, 269(16):12240–12247, Apr 1994. (Cited on page 15.)

[49] Francesca Diella, Niall Haslam, Claudia Chica, Aidan Budd, Sushama Michael, Nigel P. Brown, Gilles Trave, and Toby J. Gibson. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Frontiers in bioscience : a journal and virtual library*, 13:6580–6603, 2008. ISSN 1093-4715. URL http://view.ncbi.nlm.nih.gov/pubmed/18508681. (Cited on page 20.)

[50] M. Dimitrova, I. Imbert, M. P. Kieny, and C. Schuster. Protein-protein interactions between hepatitis C virus nonstructural proteins. *J. Virol.*, 77(9):5401–5414, May 2003. (Cited on pages 29 and 31.)

[51] H. Dinkel, S. Michael, R. J. Weatheritt, N. E. Davey, K. Van Roey, B. Altenberg, G. Toedt, B. Uyar, M. Seiler, A. Budd, L. Jodicke, M. A. Dammert, C. Schroeter, M. Hammer, T. Schmidt,

P. Jehl, C. McGuigan, M. Dymecka, C. Chica, K. Luck, A. Via, A. Chatr-Aryamontri, N. Haslam, G. Grebnev, R. J. Edwards, M. O. Steinmetz, H. Meiselbach, F. Diella, and T. J. Gibson. ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res.*, 40(Database issue):D242–251, Jan 2012. (Cited on pages 7 and 20.)

[52] H. Dinkel, K. Van Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, T. Speck, D. Kruger, G. Grebnev, M. Kuban, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, M. Seiler, L. B. Chemes, J. Glavina, I. E. Sanchez, F. Diella, and T. J. Gibson. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.*, 42(Database issue):D259–266, Jan 2014. (Cited on page 36.)

[53] A. Dömling, R. Mannhold, H. Kubinyi, and G. Folkers. *Protein-Protein Interactions in Drug Discovery*. Methods and Principles in Medicinal Chemistry. Wiley, 2013. ISBN 9783527648221. URL http://books.google.com/books?id=ySKzq-XLl5kC. (Cited on page 2.)

[54] 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. (Cited on page 3.)

[55] J. M. Doolittle and S. M. Gomez. Mapping protein interactions between Dengue virus and its human and insect hosts. *PLoS Negl Trop Dis*, 5(2):e954, 2011. (Cited on page 29.)

[56] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, 347(4):827–839, Apr 2005. (Cited on pages 21, 23, and 36.)

[57] Z. Dosztanyi, B. Meszaros, and I. Simon. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief. Bioinformatics*, 11(2):225–243, Mar 2010. (Cited on pages 21 and 36.)

[58] M. E. Droniou-Bonzom and P. M. Cannon. A systems biology starter kit for arenaviruses. *Viruses*, 4(12):3625–3646, Dec 2012. (Cited on page 31.)

[59] S. Durmuş, T. Çakir, A. Özgür, and R. Guthke. A review on computational systems biology of pathogen-host interactions. *Front Microbiol*, 6:235, 2015. (Cited on page 2.)

[60] S. D. Durmuş Tekir and K. O. Ülgen. Systems biology of pathogen-host interaction: networks of protein-protein interaction within pathogens and pathogen-human interactions in the

post-genomic era. *Biotechnol J*, 8(1):85–96, Jan 2013. (Cited on page 2.)

[61] Saliha Durmuş Tekir, Tunahan Çakır, Emre Ardıç, Ali Semih Sayılırbaş, Gökhan Konuk, Mithat Konuk, Hasret Sarıyer, Azat Uğurlu, İlknur Karadeniz, Arzucan Özgür, Fatih Erdoğan Sevilgen, and Kutlu Ö Ülgen. PHISTO: pathogen-host interaction search tool. *Bioinformatics*, 29(10):1357–1358, May 2013. (Cited on pages 25 and 26.)

[62] N. Duro, M. Miskei, and M. Fuxreiter. Fuzziness endows viral motif-mimicry. *Mol Biosyst*, 11(10):2821–2829, Sep 2015. (Cited on page 71.)

[63] M. D. Dyer, T. M. Murali, and B. W. Sobral. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics*, 23(13):i159–166, Jul 2007. (Cited on pages 3 and 71.)

[64] 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, Feb 2008. (Cited on page 32.)

[65] M. D. Dyer, T. M. Murali, and B. W. Sobral. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect. Genet. Evol.*, 11(5):917–923, Jul 2011. (Cited on page 3.)

[66] N. C. Elde and H. S. Malik. The evolutionary conundrum of pathogen mimicry. *Nat. Rev. Microbiol.*, 7(11):787–797, Nov 2009. (Cited on page 2.)

[67] E. Emmott, D. Munday, E. Bickerton, P. Britton, M. A. Rodgers, A. Whitehouse, E. M. Zhou, and J. A. Hiscox. The cellular interactome of the coronavirus infectious bronchitis virus nucleocapsid protein and functional implications for virus biology. *J. Virol.*, 87(17):9486–9500, Sep 2013. (Cited on page 31.)

[68] P. Evans, W. Dampier, L. Ungar, and A. Tozeren. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med Genomics*, 2:27, 2009. (Cited on pages 4, 7, and 52.)

[69] P. W. Ewald and H. A. Swain Ewald. Infection, mutation, and cancer evolution. *J. Mol. Med.*, 90(5):535–541, May 2012. (Cited on page 32.)

[70] S. Falkow. Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.*, 10 Suppl 2:S274–276, 1988. (Cited on page 12.)

[71] J. Filee, N. Pouget, and M. Chandler. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol. Biol.*, 8:320, 2008. (Cited on page 11.)

[72] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, 44(D1):D279–285, Jan 2016. (Cited on pages 7 and 41.)

[73] M. Flajolet, G. Rotondo, L. Daviet, F. Bergametti, G. Inchauspe, P. Tiollais, C. Transy, and P. Legrain. A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene*, 242(1-2):369–379, Jan 2000. (Cited on pages 29 and 31.)

[74] S.J. Flint. *Principles of Virology*. American Society for Microbiology, 2009. ISBN 9781555814434. URL https://books.google.com.co/books?id=W6KMmgEACAAJ. (Cited on page 13.)

[75] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010. ISSN 0370-1573. doi: DOI:10.1016/j.physrep.2009.11.002. URL http://www.sciencedirect.com/science/article/pii/S0370157309002841. (Cited on page 28.)

[76] E. Fossum, C. C. Friedel, S. V. Rajagopala, B. Titz, A. Baiker, T. Schmidt, T. Kraus, T. Stellberger, C. Rutenberg, S. Suthram, S. Bandyopadhyay, D. Rose, A. von Brunn, M. Uhlmann, C. Zeretzke, Y. A. Dong, H. Boulet, M. Koegl, S. M. Bailer, U. Koszinowski, T. Ideker, P. Uetz, R. Zimmer, and J. Haas. Evolutionarily conserved herpesviral protein interaction networks. *PLoS Pathog.*, 5(9):e1000570, Sep 2009. (Cited on page 31.)

[77] E. A. Franzosa and Y. Xia. Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.*, 108(26):10538–10543, Jun 2011. (Cited on page 32.)

[78] D. N. Fredricks and D. A. Relman. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin. Microbiol. Rev.*, 9(1):18–33, Jan 1996. (Cited on page 12.)

[79] W. Fu, B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt, and R. G. Ptak. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res.*, 37(Database issue):D417–422, Jan 2009. (Cited on pages 4, 17, 25, 26, 31, 35, 36, 37, and 47.)

[80] M. Fuxreiter, P. Tompa, and I. Simon. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, 23(8):950–956, Apr 2007. (Cited on pages 23 and 32.)

[81] T. K. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J. D. Boeke, G. Parmigiani, J. Schultz, J. S. Bader, and A. Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.*, 38(3):285–293, Mar 2006. (Cited on page 31.)

[82] J. J. Gillespie, A. R. Wattam, S. A. Cammer, J. L. Gabbard, M. P. Shukla, O. Dalay, T. Driscoll, D. Hix, S. P. Mane, C. Mao, E. K. Nordberg, M. Scott, J. R. Schulman, E. E. Snyder, D. E. Sullivan, C. Wang, A. Warren, K. P. Williams, T. Xue, H. S. Yoo, C. Zhang, Y. Zhang, R. Will, R. W. Kenyon, and B. W. Sobral. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.*, 79(11):4286–4298, Nov 2011. (Cited on pages 25 and 26.)

[83] J. Goll, S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb, and P. Uetz. MPIDB: the microbial protein interaction database. *Bioinformatics*, 24(15):1743–1744, Aug 2008. (Cited on pages 25 and 26.)

[84] M. E. Gonzalez. Vpu Protein: The Viroporin Encoded by HIV-1. *Viruses*, 7(8):4352–4368, 2015. (Cited on page 16.)

[85] J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky. Introducing protein intrinsic disorder. *Chem. Rev.*, 114(13):6561–6588, Jul 2014. (Cited on pages 20 and 21.)

[86] T. Hagai, A. Azia, A. Toth-Petroczy, and Y. Levy. Intrinsic disorder in ubiquitination substrates. *J. Mol. Biol.*, 412(3):319–324, Sep 2011. (Cited on pages 23 and 36.)

[87] T. Hagai, A. Azia, M. M. Babu, and R. Andino. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell Rep*, 7(5):1729–1739, Jun 2014. (Cited on pages 3, 4, 7, 23, 45, 46, and 52.)

[88] R. R. Halehalli and H. A. Nagarajaram. Molecular principles of human virus protein-protein interactions. *Bioinformatics*, 31(7): 1025–1033, Apr 2015. (Cited on pages 3 and 25.)

[89] S. E. Hasnain, R. Begum, K. V. Ramaiah, S. Sahdev, E. M. Shajil, T. K. Taneja, M. Mohan, M. Athar, N. K. Sah, and M. Krishnaveni. Host-pathogen interactions during apoptosis. *J. Biosci.*, 28(3):349–358, Apr 2003. (Cited on pages 1 and 72.)

[90] J. Hayer, F. Jadeau, G. Deleage, A. Kay, F. Zoulim, and C. Combet. HBVdb: a knowledge database for Hepatitis B Virus. *Nucleic Acids Res.*, 41(Database issue):D566–570, Jan 2013. (Cited on pages 17, 25, 26, and 31.)

[91] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–240, 2002. (Cited on page 28.)

[92] Z. Itzhaki. Domain-domain interactions underlying herpesvirus-human protein-protein interaction networks. *PLoS ONE*, 6(7):e21724, 2011. (Cited on page 31.)

[93] S. Jager, N. Gulbahce, P. Cimermancic, J. Kane, N. He, S. Chou, I. D'Orso, J. Fernandes, G. Jang, A. D. Frankel, T. Alber, Q. Zhou, and N. J. Krogan. Purification and characterization of HIV-human protein complexes. *Methods*, 53(1):13–19, Jan 2011. (Cited on page 31.)

[94] S. Jager, P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G. M. Jang, S. L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A. J. O'Donoghue, A. Todorovic, J. H. Morris, D. A. Maltby, T. Alber, G. Cagney, F. D. Bushman, J. A. Young, S. K. Chanda, W. I. Sundquist, T. Kortemme, R. D. Hernandez, C. S. Craik, A. Burlingame, A. Sali, A. D. Frankel, and N. J. Krogan. Global landscape of HIV-human protein complexes. *Nature*, 481(7381):365–370, Jan 2012. (Cited on page 31.)

[95] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001. (Cited on page 31.)

[96] Jeong H., Tombor B., Albert R., Oltvai Z. N., and Barabasi A.-L. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, oct 2000. ISSN 0028-0836. doi: http://dx.doi.org/10.1038/35036627. URL http://www.nature.com/nature/journal/v407/n6804/suppinfo/407651a0_S1.html. 10.1038/35036627. (Cited on page 31.)

[97] B.H. Junker and F. Schreiber. *Analysis of Biological Networks*. Wiley Series in Bioinformatics. Wiley, 2011. ISBN 9781118209912. URL https://books.google.com.co/books?id=YeXLbClh1SIC. (Cited on page 24.)

[98] K. Kadaveru, J. Vyas, and M. R. Schiller. Viral infection and human disease–insights from minimotifs. *Front. Biosci.*, 13:6455–6471, 2008. (Cited on page 23.)

[99] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, Nov 2013. (Cited on page 28.)

[100] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40(Database issue):D841–846, Jan 2012. (Cited on pages 25 and 26.)

[101] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37(Database issue):D767–772, Jan 2009. (Cited on pages 25 and 26.)

[102] S. Khadka, A. D. Vangeloff, C. Zhang, P. Siddavatam, N. S. Heaton, L. Wang, R. Sengupta, S. Sahasrabudhe, G. Randall, M. Gribskov, R. J. Kuhn, R. Perera, and D. J. LaCount. A physical interaction network of dengue virus and human proteins. *Mol. Cell Proteomics*, 10(12):M111.012187, Dec 2011. (Cited on pages 29 and 31.)

[103] C. C. Khor and M. L. Hibberd. Revealing the molecular signatures of host-pathogen interactions. *Genome Biol.*, 12(10):229, 2011. (Cited on page 19.)

[104] C. C. Khor and M. L. Hibberd. Host-pathogen interactions revealed by human genome-wide surveys. *Trends Genet.*, 28(5): 233–243, May 2012. (Cited on page 19.)

[105] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K.

Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014. (Cited on page 38.)

[106] W. Kim, M. Li, J. Wang, and Y. Pan. Biological network motif detection and evaluation. *BMC Syst Biol*, 5 Suppl 3:S5, 2011. (Cited on page 27.)

[107] M. I. Klapa, K. Tsafou, E. Theodoridis, A. Tsakalidis, and N. K. Moschonas. Reconstruction of the experimentally supported human protein interactome: what can we learn? *BMC Syst Biol*, 7:96, 2013. (Cited on page 31.)

[108] E. V. Koonin and V. V. Dolja. Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol. Mol. Biol. Rev.*, 78(2):278–303, Jun 2014. (Cited on page 11.)

[109] E. V. Koonin, T. G. Senkevich, and V. V. Dolja. The ancient Virus World and evolution of cells. *Biol. Direct*, 1:29, 2006. (Cited on page 11.)

[110] M. Kshirsagar, J. Carbonell, and J. Klein-Seetharaman. Techniques to cope with missing data in host-pathogen protein interaction prediction. *Bioinformatics*, 28(18):i466–i472, Sep 2012. (Cited on page 3.)

[111] C. Kuiken, B. Korber, and R. W. Shafer. HIV sequence databases. *AIDS Rev*, 5(1):52–61, 2003. (Cited on pages 4, 17, 35, and 36.)

[112] R. Kumar and B. Nanduri. HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinformatics*, 11 Suppl 6:S16, 2010. (Cited on pages 25 and 26.)

[113] V. Kumar, N. Kato, Y. Urabe, A. Takahashi, R. Muroyama, N. Hosono, M. Otsuka, R. Tateishi, M. Omata, H. Nakagawa, K. Koike, N. Kamatani, M. Kubo, Y. Nakamura, and K. Matsuda. Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. *Nat. Genet.*, 43(5):455–458, May 2011. (Cited on page 19.)

[114] S. K. Kwofie, U. Schaefer, V. S. Sundararajan, V. B. Bajic, and A. Christoffels. HCVpro: hepatitis C virus protein interaction database. *Infect. Genet. Evol.*, 11(8):1971–1977, Dec 2011. (Cited on pages 17, 25, 26, and 31.)

[115] N. S. Latysheva, T. Flock, R. J. Weatheritt, S. Chavali, and M. M. Babu. How do disordered regions achieve comparable functions to structured domains? *Protein Sci.*, 24(6):909–922, Jun 2015. (Cited on page 21.)

[116] J. H. Lee, V. Vittone, E. Diefenbach, A. L. Cunningham, and R. J. Diefenbach. Identification of structural protein-protein interactions of herpes simplex virus type 1. *Virology*, 378(2):347–354, Sep 2008. (Cited on page 31.)

[117] B. Lehner and A. G. Fraser. A first-draft human protein-interaction map. *Genome Biol.*, 5(9):R63, 2004. (Cited on page 31.)

[118] T. Lengauer and T. Sing. Bioinformatics-assisted anti-HIV therapy. *Nat. Rev. Microbiol.*, 4(10):790–797, Oct 2006. (Cited on pages 12 and 14.)

[119] Fei Li, Yuxing Peng, Wenjian Xu, Guangchuang Yu, Peng Li, Xiaochen Bo, and Shengqi Wang. Towards a comprehensive hbv-human interaction map. In *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS '09. International Joint Conference on*, pages 311–314, 2009. doi: 10.1109/IJCBS.2009.49. (Cited on page 31.)

[120] Q. Li, A. L. Brass, A. Ng, Z. Hu, R. J. Xavier, T. J. Liang, and S. J. Elledge. A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc. Natl. Acad. Sci. U.S.A.*, 106(38):16410–16415, Sep 2009. (Cited on pages 29 and 31.)

[121] Z. Liang, R. Liu, Y. Lin, C. Liang, J. Tan, and W. Qiao. HIV-1 Vpr protein activates the NF-kB pathway to promote G2/M cell cycle arrest. *Virol Sin*, 30(6):441–448, Dec 2015. (Cited on page 16.)

[122] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, 40(Database issue):D857–861, Jan 2012. (Cited on pages 25 and 26.)

[123] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell. Protein disorder prediction: implications for structural proteomics. *Structure*, 11(11):1453–1459, Nov 2003. (Cited on page 32.)

[124] P. Ludin, D. Nilsson, and P. Maser. Genome-wide identification of molecular mimicry candidates in parasites. *PLoS ONE*, 6(3): e17546, 2011. (Cited on page 2.)

[125] Y. Ma-Lauer, J. Lei, R. Hilgenfeld, and A. von Brunn. Virus-host interactomes–antiviral drug discovery. *Curr Opin Virol*, 2 (5):614–621, Oct 2012. (Cited on page 2.)

[126] J. P. Mackay, M. Sunde, J. A. Lowry, M. Crossley, and J. M. Matthews. Protein interactions: is seeing believing? *Trends Biochem. Sci.*, 32(12):530–531, Dec 2007. (Cited on page 24.)

[127] J. I. MacPherson, J. E. Dickerson, J. W. Pinney, and D. L. Robertson. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput. Biol.*, 6(7):e1000863, 2010. (Cited on page 17.)

[128] D. Mairiang, H. Zhang, A. Sodja, T. Murali, P. Suriyaphol, P. Malasit, T. Limjindaporn, and R. L. Finley. Identification of new protein interactions between dengue fever virus and its hosts, human and mosquito. *PLoS ONE*, 8(1):e53535, 2013. (Cited on page 31.)

[129] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U.S.A.*, 100 (21):11980–11985, Oct 2003. (Cited on page 27.)

[130] F. M. McCarthy, T. J. Mahony, M. S. Parcells, and S. C. Burgess. Understanding animal viruses using the Gene Ontology. *Trends Microbiol.*, 17(7):328–335, Jul 2009. (Cited on page 28.)

[131] A. Mehle, E. R. Thomas, K. S. Rajendran, and D. Gabuzda. A zinc-binding region in Vif binds Cul5 and determines cullin selection. *J. Biol. Chem.*, 281(25):17259–17265, Jun 2006. (Cited on page 16.)

[132] B. Meszaros, Z. Dosztanyi, and I. Simon. Disordered binding regions and linear motifs–bridging the gap between two models of molecular recognition. *PLoS ONE*, 7(10):e46829, 2012. (Cited on page 71.)

[133] L. Meyniel-Schicklin, B. de Chassey, P. Andre, and V. Lotteau. Viruses and interactomes in translation. *Mol. Cell Proteomics*, 11 (7):M111.014738, Jul 2012. (Cited on pages 23 and 32.)

[134] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002. (Cited on page 27.)

[135] M. Moerdyk-Schauwecker, D. Destephanis, E. Hastie, and V. Z. Grdzelishvili. Detecting protein-protein interactions in vesicular stomatitis virus using a cytoplasmic yeast two hybrid system. *J. Virol. Methods*, 173(2):203–212, May 2011. (Cited on page 31.)

[136] J. L. Mokili, F. Rohwer, and B. E. Dutilh. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol*, 2(1): 63–77, Feb 2012. (Cited on page 12.)

[137] R. Mosca, A. Ceol, A. Stein, R. Olivella, and P. Aloy. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, Sep 2013. (Cited on pages 25 and 26.)

[138] T. Murakami. Roles of the interactions between Env and Gag proteins in the HIV-1 replication cycle. *Microbiol. Immunol.*, 52 (5):287–295, May 2008. (Cited on page 15.)

[139] A. Nakao, M. Yoshihama, and N. Kenmochi. RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res.*, 32(Database issue):D168–170, Jan 2004. (Cited on page 41.)

[140] V. Navratil, B. de Chassey, L. Meyniel, S. Delmotte, C. Gautier, P. Andre, V. Lotteau, and C. Rabourdin-Combe. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res.*, 37(Database issue):D661–668, Jan 2009. (Cited on pages 17, 25, and 26.)

[141] V. Navratil, B. de Chassey, L. Meyniel, F. Pradezynski, P. Andre, C. Rabourdin-Combe, and V. Lotteau. System-level comparison of protein-protein interactions between viruses and the human type I interferon system network. *J. Proteome Res.*, 9(7):3527–3536, Jul 2010. (Cited on page 72.)

[142] V. Navratil, B. de Chassey, C. R. Combe, and V. Lotteau. When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC Syst Biol*, 5:13, 2011. (Cited on page 32.)

[143] V. Neduva and R. B. Russell. Linear motifs: evolutionary interaction switches. *FEBS Lett.*, 579(15):3342–3345, Jun 2005. (Cited on page 4.)

[144] NetworkX. Webpage, 2016. URL `https://networkx.github.io/`. [Online; accessed 14-January-2016]. (Cited on page 37.)

[145] M. E. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Nov 2002. (Cited on page 31.)

[146] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103 (23):8577–8582, 2006. doi: 10.1073/pnas.0601602103. URL http://www.pnas.org/content/103/23/8577.abstract. (Cited on page 28.)

[147] Newman M. E. J. Communities, modules and large-scale structure in networks. *Nat Phys*, 8(1):25–31, jan 2012. ISSN 1745-2473. doi: http://dx.doi.org/10.1038/nphys2162. 10.1038/nphys2162. (Cited on page 28.)

[148] A. Ng, B. Bursteinas, Q. Gao, E. Mollison, and M. Zvelebil. Resources for integrative systems biology: from data through databases to networks and dynamic system models. *Brief. Bioinformatics*, 7(4):318–330, Dec 2006. (Cited on page 25.)

[149] E. Nourani, F. Khunjush, and S. Durmu? Computational approaches for prediction of pathogen-host protein-protein interactions. *Front Microbiol*, 6:94, 2015. (Cited on page 3.)

[150] I. Nouretdinov, A. Gammerman, Y. Qi, and J. Klein-Seetharaman. Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. *Pac Symp Biocomput*, pages 311–322, 2012. (Cited on page 48.)

[151] M. B. Oldstone. Molecular mimicry and immune-mediated diseases. *FASEB J.*, 12(13):1255–1265, Oct 1998. (Cited on page 1.)

[152] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman, F. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. Hancock, R. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stumpflen, M. Tyers, P. Uetz, I. Xenarios, and H. Hermjakob. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, 9(4):345–350, Apr 2012. (Cited on page 25.)

[153] L. Padilla-Noriega, O. Paniagua, and S. Guzman-Leon. Rotavirus protein NSP3 shuts off host cell protein synthesis. *Virology*, 298(1):1–7, Jun 2002. (Cited on page 39.)

[154] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H. W. Mewes, A. Ruepp, and D. Frishman. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21 (6):832–834, Mar 2005. (Cited on page 26.)

[155] J. Pan, X. Peng, Y. Gao, Z. Li, X. Lu, Y. Chen, M. Ishaq, D. Liu, M. L. Dediego, L. Enjuanes, and D. Guo. Genome-wide analysis of protein-protein interactions and involvement of viral proteins in SARS-CoV replication. *PLoS ONE*, 3(10):e3299, 2008. (Cited on page 31.)

[156] S. C. Pettit, S. Gulnik, L. Everitt, and A. H. Kaplan. The dimer interfaces of protease and extra-protease domains influence the activation of protease and the specificity of GagPol cleavage. *J. Virol.*, 77(1):366–374, Jan 2003. (Cited on pages 14 and 36.)

[157] E. M. Phizicky and S. Fields. Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.*, 59(1):94–123, Mar 1995. (Cited on page 24.)

[158] J. W. Pinney, J. E. Dickerson, W. Fu, B. E. Sanders-Beer, R. G. Ptak, and D. L. Robertson. HIV-host interactions: a map of viral perturbation of the host system. *AIDS*, 23(5):549–554, Mar 2009. (Cited on page 17.)

[159] C. P. Ponting and R. R. Russell. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71, 2002. (Cited on page 20.)

[160] C. Prieto and J. De Las Rivas. APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res.*, 34(Web Server issue):298–302, Jul 2006. (Cited on pages 25 and 26.)

[161] N. Przulj and D. J. Higham. Modelling protein-protein interaction networks via a stickiness index. *J R Soc Interface*, 3(10): 711–716, Oct 2006. (Cited on pages 30 and 32.)

[162] R. G. Ptak, W. Fu, B. E. Sanders-Beer, J. E. Dickerson, J. W. Pinney, D. L. Robertson, M. N. Rozanov, K. S. Katz, D. R. Maglott, K. D. Pruitt, and C. W. Dieffenbach. Cataloguing the HIV type 1 human protein interaction network. *AIDS Res. Hum. Retroviruses*, 24(12):1497–1502, Dec 2008. (Cited on page 17.)

[163] R. Pushker, C. Mooney, N. E. Davey, J. M. Jacque, and D. C. Shields. Marked variability in the extent of protein disorder within and between viral families. *PLoS ONE*, 8(4):e60724, 2013. (Cited on page 23.)

[164] 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, May 2006. (Cited on page 3.)

[165] X. Qian and B. J. Yoon. Comparative analysis of protein interaction networks reveals that conserved pathways are susceptible

to HIV-1 interception. *BMC Bioinformatics*, 12 Suppl 1:S19, 2011. (Cited on page 17.)

[166] Tali H. Reingewertz, Deborah E. Shalev, and Assaf Friedler. *Making Order in the Intrinsically Disordered Regions of HIV-1 Vif Protein*, pages 201–221. John Wiley & Sons, Inc., 2011. ISBN 9781118135570. doi: 10.1002/9781118135570.ch8. URL http://dx.doi.org/10.1002/9781118135570.ch8. (Cited on pages 16 and 36.)

[167] S. Ren, V. N. Uversky, Z. Chen, A. K. Dunker, and Z. Obradovic. Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC Genomics*, 9 Suppl 2:S26, 2008. (Cited on page 45.)

[168] J. F. Roeth and K. L. Collins. Human immunodeficiency virus type 1 Nef: adapting to intracellular trafficking pathways. *Microbiol. Mol. Biol. Rev.*, 70(2):548–563, Jun 2006. (Cited on page 16.)

[169] P. W. Rose, C. Bi, W. F. Bluhm, C. H. Christie, D. Dimitropoulos, S. Dutta, R. K. Green, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, A. G. Ramos, J. D. Westbrook, J. Young, C. Zardecki, H. M. Berman, and P. E. Bourne. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, 41 (Database issue):D475–482, Jan 2013. (Cited on page 3.)

[170] R. Rozen, N. Sathish, Y. Li, and Y. Yuan. Virion-wide protein interactions of Kaposi's sarcoma-associated herpesvirus. *J. Virol.*, 82(10):4742–4750, May 2008. (Cited on page 31.)

[171] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005. (Cited on page 31.)

[172] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, 32(Database issue):D449–451, Jan 2004. (Cited on pages 25 and 26.)

[173] D. P. Sargeant, M. R. Gryk, M. W. Maciejewski, V. Thapar, V. Kundeti, S. Rajasekaran, P. Romero, K. Dunker, S. C. Li,

T. Kaneko, and M. R. Schiller. Secondary structure, a missing component of sequence-based minimotif definitions. *PLoS ONE*, 7(12):e49957, 2012. (Cited on page 71.)

[174] D. Sarkar, T. Jana, and S. Saha. LMPID: a manually curated database of linear motifs mediating protein-protein interactions. *Database (Oxford)*, 2015, 2015. (Cited on pages 35, 37, 47, 48, and 50.)

[175] M. Sarmady, W. Dampier, and A. Tozeren. Sequence- and interactome-based prediction of viral protein hotspots targeting host proteins: a case study for HIV Nef. *PLoS ONE*, 6(6): e20735, 2011. (Cited on page 28.)

[176] A. Segura-Cabrera, C. A. Garcia-Perez, X. Guo, and M. A. Rodriguez-Perez. A viral-human interactome based on structural motif-domain interactions captures the human infectome. *PLoS ONE*, 8(8):e71526, 2013. (Cited on pages 32 and 71.)

[177] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, Nov 2003. (Cited on page 8.)

[178] S. D. Shapira, I. Gat-Viks, B. O. Shum, A. Dricot, M. M. de Grace, L. Wu, P. B. Gupta, T. Hao, S. J. Silver, D. E. Root, D. E. Hill, A. Regev, and N. Hacohen. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*, 139(7):1255–1267, Dec 2009. (Cited on page 31.)

[179] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, 35(Database issue): D786–793, Jan 2007. (Cited on page 21.)

[180] I. Sillitoe, A. L. Cuff, B. H. Dessailly, N. L. Dawson, N. Furnham, D. Lee, J. G. Lees, T. E. Lewis, R. A. Studer, R. Rentzsch, C. Yeats, J. M. Thornton, and C. A. Orengo. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, 41 (Database issue):D490–498, Jan 2013. (Cited on page 20.)

[181] N. Simonis, J. F. Rual, I. Lemmens, M. Boxus, T. Hirozane-Kishikawa, J. S. Gatot, A. Dricot, T. Hao, D. Vertommen, S. Legros, S. Daakour, N. Klitgord, M. Martin, J. F. Willaert, F. Dequiedt, V. Navratil, M. E. Cusick, A. Burny, C. Van Lint,

D. E. Hill, J. Tavernier, R. Kettmann, M. Vidal, and J. C. Twizere. Host-pathogen interactome mapping for HTLV-1 and -2 retroviruses. *Retrovirology*, 9:26, 2012. (Cited on page 31.)

[182] A. E. Smith and A. Helenius. How viruses enter animal cells. *Science*, 304(5668):237–242, Apr 2004. (Cited on pages 16 and 72.)

[183] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.*, 100(21):12123–12128, Oct 2003. (Cited on page 28.)

[184] T. Stellberger, R. Hauser, A. Baiker, V. R. Pothineni, J. Haas, and P. Uetz. Improving the yeast two-hybrid system with permutated fusions proteins: the Varicella Zoster Virus interactome. *Proteome Sci*, 8:8, 2010. (Cited on page 31.)

[185] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, Sep 2005. (Cited on page 31.)

[186] R. J. Sugrue, C. Brown, G. Brown, J. Aitken, and H. W. McL Rixon. Furin cleavage of the respiratory syncytial virus fusion protein is not a requirement for its transport to the surface of virus-infected cells. *J. Gen. Virol.*, 82(Pt 6):1375–1386, Jun 2001. (Cited on page 20.)

[187] B. Suter, S. Kittanakom, and I. Stagljar. Two-hybrid technologies in proteomics research. *Curr. Opin. Biotechnol.*, 19(4):316–323, Aug 2008. (Cited on page 24.)

[188] N. H. Tan Gana, T. Onuki, A. F. Victoriano, and T. Okamoto. MicroRNAs in HIV-1 infection: an integration of viral and cellular interaction at the genomic level. *Front Microbiol*, 3:306, 2012. (Cited on page 12.)

[189] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman. Refining literature curated protein interactions using expert opinions. *Pac Symp Biocomput*, pages 318–329, 2015. (Cited on page 51.)

[190] D. L. Theobald. A formal test of the theory of universal common ancestry. *Nature*, 465(7295):219–222, May 2010. (Cited on page 11.)

[191] P. Tompa, N. E. Davey, T. J. Gibson, and M. M. Babu. A million peptide motifs for the molecular biologist. *Mol. Cell*, 55(2):161–169, Jul 2014. (Cited on pages 4 and 25.)

[192] C. Torresilla, J. M. Mesnard, and B. Barbeau. Reviving an old HIV-1 gene: the HIV-1 antisense protein. *Curr. HIV Res.*, 13(2): 117–124, 2015. (Cited on page 14.)

[193] D. Tortorella, B. E. Gewurz, M. H. Furman, D. J. Schust, and H. L. Ploegh. Viral subversion of the immune system. *Annu. Rev. Immunol.*, 18:861–926, 2000. (Cited on page 16.)

[194] P. Uetz, Y. A. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S. V. Rajagopala, M. Roupelieva, D. Rose, E. Fossum, and J. Haas. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239–242, Jan 2006. (Cited on page 31.)

[195] Uniprot. Ftp server, 2016. URL `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/`. [Online; accessed 14-January-2016]. (Cited on page 37.)

[196] K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, and N. E. Davey. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.*, 114(13):6733–6778, Jul 2014. (Cited on page 23.)

[197] K. Van Vliet, M. R. Mohamed, L. Zhang, N. Y. Villa, S. J. Werden, J. Liu, and G. McFadden. Poxvirus proteomics and virus-host protein interactions. *Microbiol. Mol. Biol. Rev.*, 73(4):730–749, Dec 2009. (Cited on page 31.)

[198] V. Vittone, E. Diefenbach, D. Triffett, M. W. Douglas, A. L. Cunningham, and R. J. Diefenbach. Determination of interactions between tegument proteins of herpes simplex virus type 1. *J. Virol.*, 79(15):9566–9571, Aug 2005. (Cited on page 31.)

[199] V. A. Volchkova, H. D. Klenk, and V. E. Volchkov. Delta-peptide is the carboxy-terminal cleavage fragment of the nonstructural small glycoprotein sGP of Ebola virus. *Virology*, 265(1):164–171, Dec 1999. (Cited on page 20.)

[200] A. von Brunn, C. Teepe, J. C. Simpson, R. Pepperkok, C. C. Friedel, R. Zimmer, R. Roberts, R. Baric, and J. Haas. Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS ONE*, 2(5):e459, 2007. (Cited on page 31.)

[201] D. Walsh and I. Mohr. Viral subversion of the host protein synthesis machinery. *Nat. Rev. Microbiol.*, 9(12):860–875, Dec 2011. (Cited on pages 7, 16, 39, and 41.)

[202] B. M. Ward. The taking of the cytoskeleton one two three: how viruses utilize the cytoskeleton during egress. *Virology*, 411(2): 244–250, Mar 2011. (Cited on pages 16 and 72.)

[203] T. Watanabe, S. Watanabe, and Y. Kawaoka. Cellular networks involved in the influenza virus life cycle. *Cell Host Microbe*, 7(6): 427–439, Jun 2010. (Cited on page 31.)

[204] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998. (Cited on page 31.)

[205] J. M. Watts, K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess, R. Swanstrom, C. L. Burch, and K. M. Weeks. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460(7256):711–716, Aug 2009. (Cited on page 13.)

[206] R. Winnenburg, M. Urban, A. Beacham, T. K. Baldwin, S. Holland, M. Lindeberg, H. Hansen, C. Rawlings, K. E. Hammond-Kosack, and J. Kohler. PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res.*, 36 (Database issue):D572–576, Jan 2008. (Cited on pages 25 and 26.)

[207] W. Wu, K. C. Tran, M. N. Teng, K. J. Heesom, D. A. Matthews, J. N. Barr, and J. A. Hiscox. The interactome of the human respiratory syncytial virus NS1 protein highlights multiple effects on host cell biology. *J. Virol.*, 86(15):7777–7789, Aug 2012. (Cited on page 31.)

[208] S. Wuchty. Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens. *PLoS ONE*, 6(11):e26960, 2011. (Cited on page 3.)

[209] S. Wuchty, G. Siwo, and M. T. Ferdig. Viral organization of human proteins. *PLoS ONE*, 5(8):e11796, 2010. (Cited on pages 28 and 32.)

[210] B. Xue, R. W. Williams, C. J. Oldfield, G. K. Goh, A. K. Dunker, and V. N. Uversky. Viral disorder or disordered viruses: do viral proteins possess unique features? *Protein Pept. Lett.*, 17(8): 932–951, Aug 2010. (Cited on page 21.)

[211] B. Xue, M. J. Mizianty, L. Kurgan, and V. N. Uversky. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell. Mol. Life Sci.*, 69(8):1211–1259, Apr 2012. (Cited on pages 17 and 45.)

[212] L. Zhang, N. Y. Villa, M. M. Rahman, S. Smallwood, D. Shattuck, C. Neff, M. Dufford, J. S. Lanchbury, J. Labaer, and G. McFadden. Analysis of vaccinia virus-host protein-protein interactions: validations of yeast two-hybrid screenings. *J. Proteome Res.*, 8(9):4311–4318, Sep 2009. (Cited on page 31.)

[213] Z. Zhou, L. J. Licklider, S. P. Gygi, and R. Reed. Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419 (6903):182–185, Sep 2002. (Cited on page 28.)

[214] R. Zoraghi and N. E. Reiner. Protein interaction networks as starting points to identify novel antimicrobial drug targets. *Curr. Opin. Microbiol.*, Aug 2013. (Cited on page 2.)

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. `classicthesis` is available for both LaTeX and LyX:

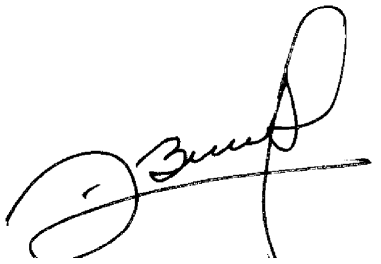
http://code.google.com/p/classicthesis/

Final Version as of September 12, 2016 (`classicthesis` final version).

DECLARATION

This thesis is submitted to opt for the title of *Doctor en ingeniería, énfasis en ciencias de la computación* in the Escuela de Ingeniería de Sistemas y Computación in the Facultad de Ingeniería of the Universidad del Valle.

Santiago de Cali, September - 2016



Andrés Becerra Sandoval