Abstract

Viral proteins rely extensively on the molecular mimicry of cellular linear motifs for modifying cell signalling and other processes in ways that favour viral infection. This study aims to de-novo discover human linear motifs convergently evolved in disordered regions of viral proteins. We systematically apply computational motif prediction tools to the human-viral and human-human protein interaction network. By limiting the search space to the sequences of viral proteins we can increase the sensitivity of motif prediction. We can recover known linear motif occurrences in viral proteins. By estimating recognition domain identities and utilising alternative human-human interaction datasets we further improve our ability to recover these motifs. We plan to experimentally validate novel human recognition domain - viral motif pairs using phage display. This work contributes to our understanding of the domain-linear motif code and how viruses exploit this mechanism.

This Master's thesis contains XX pages, is illustrated by XX figures and X tables. Number of references is XX.

Keywords: short linear motifs prediction, protein interaction networks, convergent evolution, molecular mimicry, viral proteins

Introduction

Linear motifs are short amino-acid sequence motifs that mediate physical and selective protein-protein interactions. Linear motifs are usually located in the disordered regions of a protein and are usually recognized by structured globular domains [24926813].

Linear motif-mediated interactions are known to connect and direct cell signaling pathways [PMC3664230]. This function is often further regulated by post-translational modifications and cooperativity [PMC4666095]. Linear motif-mediated interactions can evolve rapidly and help rewire cell signaling networks during speciation events, in disease or in host-pathogen interactions [27540857, PMC4089993, PMC4654906].

A number of linear motifs have been identified using traditional molecular biology approaches and hypothesis-driven research, however, those methods are laborious and most of the functional linear motifs are yet to be identified [24926813]. Using computational motif search tools to identify linear motifs in homologous proteins tend to result in a large number of non-functional. A number of approaches have been shown to improve the efficiency of identifying functional motifs: incorporating protein-protein interaction data, the sequence conservation across species and filtering for motifs located in the unstructured regions [25555723, PMC4652402]. Methods such as phage display were developed to aid experimental discovery of motifs at a proteome scale [26297553]. However, we are far from full characterization of the domain-linear motif code and current estimate suggest that only 1% of motif instances have been discovered to date compared to 15-40% of estimated interactions [15943979, 16962311].

The aim of this study is to use host-viral protein interactions data as a way to limit the search space to identify novel functional linear motifs. Viral proteins mimic cellular linear motifs to interact with and modify cell signaling in a way that favours the progression of viral infection [PMC4089993]. We can use this functional relationship to improve the sensitivity of computational motif prediction. This analysis had not been done before with such a large dataset as well as using a

combination of viral-human and human networks. Unlike other interactome-wide studies [21879107], we also use a statistical method to estimate which domains are likely to mediate interaction. Computationally predicted domain-linear motif pairs will be verified using phage display screens in a collaborating laboratory.

This work can contribute to the understanding of the domain-linear motif code and how viruses exploit this mechanism.

The tasks of this project were:

1. Retrieve and process experimental interaction data from public databases and examine the properties of viral-human interaction network.

2. Use viral-human network, probabilistic motif search tools to de-novo discover short linear motifs. Use the sequences of viral proteins to limit the search space.

3. Identify protein sequence domains in all viral and human proteins. Estimate which human domains are likely to mediate interaction with each viral protein.

4. Evaluate our motif search pipeline against a benchmark of known viral motifs.

5. Implement this motif search pipeline in R statistical programming language, using command-line tools and LSF high-performance computing cluster.

1. Literature Review

1.1 Short linear motifs

1.1.1 Protein interactions modules

Commented [EP1]: This section is a bit hard to read, slightly repeptitive and lacks references.

The structure and functions of cells arise from interactions between molecules inside and outside it [@Hein:2015aa]. Proteins, nucleic acids, lipids and small molecules can all form biologically important interactions. In our study, we focus on interactions between proteins. These interactions govern cellular processes and organismal phenotypes from cell death to muscle contraction. Disrupting or introducing a new protein contact can constitute a molecular basis of disease or an evolutionary adaptation [15993577, 27540857, 24882001]. To produce these phenotypes, proteins interact under specific conditions in defined cell types and subcellular locations [28746306]. This way interactions organise biochemical and enable structural functions of the protein.

All aspects of protein function are conducted by modules embedded in its sequence. These modules can be folded in a stable 3D structure under native conditions (globular domains) or lack a stable 3D structure (disordered regions). Globular domains reserve a variety of functions requiring a precise spatial position of amino acid residues and a rigid structure: enzymatic, ligand-binding (DNA, lipids, peptides) or structural functions. Globular domains constitute a majority of known protein interaction interfaces, however, most of these interactions are very stable and lack dynamical properties necessary for inducible and transient interactions. The functionality of globular domain-mediated interactions is complemented by short linear motifs (SLIMs or linear motifs) located in flexible disordered regions [24926813].

1.1.2 Short linear motifs and molecular recognition features

Short Linear Motifs (SLIMs) are sequence motifs of 3-15 amino acid residues that mediate physical and selective interactions between proteins. The linear

sequence of the motif and not it's three-dimensional structure is considered to be important for binding. SLiMs are usually located in the disordered part of a protein [21909575, 17387114]. This can be a long-disordered region or a short loop on the surface of a globular domain [12384576]. The flexibility of this region allows globular domains of the interacting partner to recognise the SLiM. Only 1-5 amino acid residues are required determinants of recognition specificity such as phosphotyrosine in the SH2-domain binding motif [PMC2984226]. Amino acid residues in neighbouring positions can further modify specificity, affinity and selectivity motifs. Permissive sequences enhance binding while non-permissive oppose binding close to the essential site. Which amino acids are essential, permissive or non-permissive is in most cases specific recognition domain instance. For example, [PMC2984226]. The shape and physicochemical properties of the binding pocket determines the sequence specificity, affinity and selectivity of recognition domain [24926813].

Domain families can have a broad specificity for a class of motifs. For example, SH2, SH3 and PDZ domains bind to phosphotyrosine, proline-rich or C-terminal motifs respectively. In contrast, a specific instance of a domain in a protein can recognise a more specific motif sequence in a limited set of proteins. For example, the SH2 domain of GRB2 binds phosphorylated pYENV motif of receptor tyrosine kinases leading to inducible recruitment, while the SH2 domain of Src recognises pYEEI motif in the sequence of Src itself causing the autoinhibition of the kinase (PMID: 11719057). The sequence context around phosphorylated tyrosines determines which SH2 domain-containing proteins will be recruited and which signalling pathways will be activated. Other docking motifs and their recognition domains complement the low specificity of protein-serine/threonine kinase domains (such as those of MAP kinases) towards their targets (PMID:11371562). As illustrated in these examples, SLiM-mediated interactions can be inducible, transient and serve regulatory functions. Both domains and motifs can have a different degree of binding specificity towards their partners. Multiple domains can recognise the

same promiscuous motifs or the same promiscuous domain can recognise multiple motifs [24926813].

Motif-mediated interactions are the weakest of 3 main types: domain-domain interaction, domain-motif interactions and domain - molecular recognition feature (or MORF) interactions. These types of interactions differ by the area of interaction interface that, in turn, contributes to affinity. Domain-domain interactions are the strongest (picomolar affinity) and are usually involved in stable protein complex formation. In domain-MORF interactions a disordered region of one protein undergoes disorder-to-order transition gaining a stable 3-dimensional structure in a complex. MORFs are also called intrinsically disordered domains. These interactions have the intermediate strength (nanomolar affinity). The low affinity of motif-mediated interactions (low micromolar affinity) affects the dynamics of interactions enabling fast switching or requiring the presence of multiple motifs for high-avidity biologically relevant binding. This along with other properties of SLIMs makes the ideal for connecting protein complexes (@Hein:2015aa, PMC4191887), targeting proteins to a specific organelle or assembling functionally different complexes around the invariant core (proteasome, transcription initiation machinery). This way, SLIM-mediated interactions constitute the molecular basis for almost all cellular processes.

1.1.3 Classes of Short Linear Motifs

Motifs can be classified into 2 general groups: motifs that mediate binding or motifs that are the target for post-translational modification (PTM). Each of these groups can be further subdivided. Binding-mediating SLiMs include ligand-binding, targeting, docking and degradation motifs. PTM motifs are subdivided into moiety addition/removal motifs, or classic PTM motifs, and cleavage motifs [24926813]. A total of 6 motif types are annotated in the ELM database that collects instance of known motifs from the literature [PMC5753338].

Commented [EP2]: localisation?

Commented [VK3R2]: No. Review by Roey, Gibson, Davey (https://www.ncbi.nlm.nih.gov/pubmed/24926813) classifies motif exactly into these groups: binding and PTM targets.

Commented [EP4]: This is confusing. I'd start with the two categories and include there the other 6.

Ligand-binding motifs

Classic ligand-binding motifs mediate a protein complex assembly - including functionally distinct complexes around the same invariant core or the scaffolding of proteins forming the same pathway. For example, nuclear receptors recruit transcriptional repressors or activators via CoRNR motif or NR box motif respectively depending on their binding of steroid hormone - their ligand [15276186]. Scaffold proteins can regulate cell signalling in multiple ways: from specifying linear pathways by arranging kinases the right order (such as KSR that organises MAP kinase cascade) to inhibition via scaffold titration or allosteric regulation [21551057].

Targeting motifs drive the correct localisation of proteins by directing translocation of proteins between subcellular compartments via specific transport pathways (trafficking motifs, e.g. nuclear localisation signal, nuclear export signal) - or by retaining a protein in the correct location via anchoring this protein to a compartment-specific complex (SxIP-motif recognised by EBH domain of microtubule end-binding proteins [22885064]).

Docking motifs serve to recruit modifying enzymes and are used extensively to increase the substrate specificity of enzymes. Docking motifs often guide enzymes to modify a different site in the substrate. Docking motifs can drive substrate recognition via 3 main modes. Motifs can be recognised by a site in a catalytic domain distinct from the catalytic site. For example, a docking groove of MAP kinase catalytic domain recognises a docking motif in MEF2A, MAP2K1 or MKP1. Alternatively, a separate interaction module located in the same protein recognises a docking motif. Previously mentioned SH2-domains often serve this function. Not only the SH2 domain of Src kinase is involved in autoinhibition but also in substrate recruitment such as recruiting FAK1 kinase by recognising pYAEI motif [11604500]. Finally, a recognition domain that binds a substrate can be located in a different protein (not the enzyme itself) that forms a complex with the enzyme. This complex has to be assembled first and may rely on either linear-motif or domain-domain -mediated interactions. One example is CDK (cyclin-dependent

kinases) that rely on a recognition domain in a cyclin protein to recognise their targets [16707497].

Commented [EP5]: A nice figure would be a structure with e.g. a kinase and the peptide binding to the docking motif and the catalytic site, or any other example. You can make a schematic of all these ways and add a structure of an example

A special subset of docking motifs that regulate protein stability is called degradation motifs or degrons (DEG in ELM database). These motifs dock the ubiquitin ligase enzyme (such as E3-culin complex) to their substrates. Ubiquitin tagging of these substrates target them for degradation by the proteasome - so-called ubiquitin-proteasome system. It should be noted that depending on the number of ubiquitins attached or the structure of poly-ubiquitin - this mark can control protein-protein interaction and subcellular localization in addition to degradation (K48 polyubiquitin is a degradation mark)[15571809, 20704751]. Ubiquitin ligases form cascades of E1, E2 and E3 ligases that generate activated E2-ubiquitin for tagging of substrates by E3 ligases (http://dx.doi.org/10.1016/j.cell.2016.03.003). Last stage E3 ligases are considered as main determinants of degradation specificity and are the most prevalent in human genomes (>700 E3 enzymes, ~40 E2 enzymes, 2 E1 enzymes) [15571809, 25394868]. These proteins recognise degrons often in a context-dependent and inducible manner with RING, U-box and Cullin-RING ligases requiring different regulation [http://dx.doi.org/10.1016/j.cell.2016.03.003, 15688063].

Post-translational modification motifs

The second large group of motifs overlaps with sites of post-translational modification (PTM) and mediates a recognition of the substrate by the active site of an enzyme. There are 3 main classes of post-translational modification motifs:

1. Motifs recognised by an enzyme that catalyses an addition or removal of a group, such as a phosphate, ubiquitin or lipid (MOD in ELM).

2. Motifs recognised by a cleavage enzyme (CLV).

3. Motifs recognised by an enzyme that catalyses the cis-trans conversion of a proline peptide bond. These enzymes are called peptidyl-prolyl cis-trans isomerases with the most prominent example being Cyclophilin family of proteins and PIN1 [24926813].

Many types of addition-removal of a group type PTMs have been discovered: phosphorylation, acetylation, methylation, sumoylation, ubiquitination, attachment of lipids for anchoring to the membrane and many other, less prevalent modifications [17585314, 25053359, 25491103, 28488703, 22781905, 23175280]. These are used extensively across multiple cellular processes with phosphorylation-mediated signalling and epigenetic control of gene expression being the most studied. Epigenetic control in this context describes the modification of disordered tails of histone proteins at different sites controls chromatin state and transcription [PMC3193420]. These modifications often constitute the context for ligand-binding motifs by disrupting or enabling interactions directly or via cooperative mechanisms involving charge-induced structural change, multiple motifs or binding partners [24926813]. For example, SH2-domain of GRB2 binds to a phosphotyrosine residue of tyrosine kinase receptor once it got phosphorylated [8816475]. Often, we do not have a definitive proof which mechanism is being used. For example, phosphorylation of KSR1 and Raf by activated Erk blocks Raf binding to KSR1 to change the dynamics of MAPK signalling (adaptation to the input signal). This phosphorylation event is likely to act by affecting the ability of a recognition domain in KSR1 protein to bind a proline-rich motif in Raf protein rather than by modifying motif [PMC2708738]. Another example of MAPK signalling in yeast shows that the phosphorylation of a scaffold protein Ste5 enables a recognition domain to interact with Fus3 (the last kinase in a cascade) resulting in ultrasensitive MAPK signalling dynamics [20400943]. MAPK signaling activity does not change until a receptor stimulation level has reached a certain threshold after which it quickly increases to a high activity steady state (plateau).

Commented [EP6]: I don't know what this means, too vague

Cleavage motifs are recognised by a catalytic domain of proteases (similar to modification motifs) and irreversibly hydrolysed at the cleavage site (unlike modification sites). These enzymes perform limited proteolysis disrupting or sometimes enabling the function of a protein. The most prominent motifs of this class are those recognised by caspases - the main drivers of programmed cell death or apoptosis [23545416] or the inflammation response in myeloid cells [26121197].

The apoptosis pathway can be initiated by immune cells (such as T-killers or natural killers) externally or by the damage of DNA or mitochondria internally and usually starts from binding-induced activation of regulatory caspases. Regulatory caspases (e.g. caspase 8 & 9) activate effector caspases (e.g. caspase 3, 6 & 7) by recognising LEHD motif and cleaving them [PMC3721276]. Effector caspases than cleave 100s of proteins resulting in apoptosis. Caspases 3 and 7 recognise [DSTE][^P][^DEWHFYC]D[GSAN] motif (CLV_C14_Caspase3-7 in ELM) in their substrates. In this example, a motif is specified by a regular expression that will be discussed later. By acquiring this recognition motif, a protein can get under the control of the apoptosis pathway [24926813]. Effector caspases can have both regulatory (activation of DNA cleavage enzyme) and disrupting (cleavage of cytoskeletal proteins) effects on its substrates.

All of these classes are partially overlapping and not mutually exclusive, for example, the ligand-binding site can attach a protein to a complex but also determine its subcellular localisation (targeting). The same motif may be a post-translational modification motif and a classic ligand-binding motif (SH2-domain ligand) for a. So, motif classes are defined in the context of interaction rather than as a property of the sequence. Such context-dependence and a functional definition make identifying linear motifs challenging [26581338]. In addition, as will be discussed in a later chapter, the same sequence of amino acids can be a functional motif or not a functional motif depending on whether it is accessible for binding to recognition domains.

Commented [EP7]: REFERENCES!!!

1.2 Changing function by interaction network rewiring via protein expression, evolution and splicing

To illustrate how the evolution of linear motifs can produce a new function by rewiring of the protein interaction network let's examine the example of docking motifs. As described in the previous section, docking motifs function by placing the substrate in a close proximity of a catalytic domain, that is increasing the local concentration of the substrate and, thus, allowing to achieve the specificity and

selectivity (orthogonality) of signalling response (multiple inputs - same kinase - multiple inputs-dependent substrates) via spatial separation of irrelevant enzymes and substrates. In this light, it is important to stress the dynamic and conditional nature of SLiM-mediated interactions and the quantitative nature of cell signalling reactions. While off-target substrates may be phosphorylated at some rate only correct substrates will be modified at a rate sufficient to elicit a biologically meaningful response (as a result of spatial proximity) [18339942]. Modularity of protein sequence allows putting arbitrary catalytic domains and linear motifs into this protein sequence to bring a different catalytic function to the relevant protein complex or location determined by SLIM.

The fact that the linear-motif-mediated spatial proximity (not the precise protein structure) is enough for many regulatory interactions increases the evolutionary plasticity of these interactions. If you can find a way to put an antiviral host protein, such as the cytidine deaminase APOBEC3G, in a close proximity of the Cullin-E2 ubiquitin ligase (e.g. by designing a scaffold protein containing motifs that both of them recognise) you can hijack cell's own system to enable viral infection [14564014]. This is an example of network rewiring via protein expression - in this case of a viral protein, but the same mechanism can control the functional diversity of cell types by using proteins with cell-lineage-restricted expression.

Another process relying on spatial proximity is gene regulation. Cells can activate the transcription of oncogenes by fusing a DNA-binding domain that binds to promoters of these genes (FLI1 protein) to a disordered region containing motifs that recruit transcription activation machinery (trans-activation domain, EWSR1 proteins) [27540857]. Such gene fusion events in cancer tend to disrupt interactions of proteins with other molecules: proteins, RNA, DNA. Disordered protein regions containing linear motifs and posttranslational modifications sites may be selectively excluded in the fusion protein lifting off regulatory control. This is an example of network rewiring by mutagenic events with subsequent phenotypic selection (in this case for ability to proliferate uncontrollably). While this is an example of linear

motif evolution in a human disease, cancer, analogous processes can act to change cell functions on a larger evolutionary timescale [18753782].

Our understanding of linear motif role in functional innovation has improved by analysing interactions between proteins with distinct annotated functions and how these interactions evolve along the phylogenetic tree [25299147]. We discuss this in the next paragraph.

In their study Kim et al define data-driven protein interaction modules that, in theory, correspond to protein complexes. SLIM-mediated interactions are more likely to connect proteins from modules with different function while domain-domain interactions are more likely to connect proteins within modules. SLIM- or domain-mediated interactions were predicted. Function of modules was determined using Gene Ontology annotation of genes in a module. Kim et al also showed that complex metazoan species gained more motif-mediated interaction than domain-mediated interactions [25299147]. In an independent study, Hein et al overplayed experimentally determined interaction affinity on the network topology. They found, that proteins within modules are connected by strong interactions while protein contacts between modules were weak. While Hein et al did not demonstrate that these weak interactions are SLIM-mediated given what we know about affinity of these interactions it would be fair to hypothesise that.

Finally, interactions can be rewired by splicing motifs in and out to alter protein localisation or targetability by enzymes [22749400].

1.3 Linear motif evolution

To examine evolutionary properties of linear motifs it is best to contrast them to globular domains. Mechanisms of domain evolution are pretty much a textbook knowledge while a complete understanding of motif evolution is still being established. Domains evolve via duplication, divergence and recombination [21286315]. In contrast, SLIMs often evolve de-novo, or ex-nihilo, in sequences of both non-homologous or homologous proteins. Non-homologous proteins can gain the same motif. Homologous protein can evolve new motif classes not shared by

their common ancestor [26589632]. Homologous proteins may lose a motif they shared and instead gain the same motif in a sequence of the same disordered region. This phenomenon is called a turnover of motifs. To better understand these phenomena, it is necessary to consider the sequence and structural context in which motifs evolve.

1.3.1 Disordered regions provide context

The sequence of disordered regions is not constrained by structural context allowing rapid sequence evolution. Amino acid substitutions do not have a disrupting effect on the protein structure if they happen in a disordered region leading to lower cost multiple sequential substitutions. Disordered regions provide a context in which a short linear motif can evolve in a few or even one substitution event [24773235, 24926813]. This context provides the conditions necessary for convergent evolution of SLIMs.

1.3.2 Selective forces and convergent evolution of SLIMs

Next, let's examine how selective forces act on ex-nihilo evolving motifs. If a new motif is never recognised in the right context and does not provide an evolutionary advantage or disadvantage this motif will be lost via the same process of random mutation. Positive selection will keep a motif that offers useful modes of regulation. A model that under which 2 proteins can independently gain certain sequence is called convergent. The prevalent strategy for evolving short linear motif is convergence (in contrast to domain structure and domain architecture of proteins). This is exemplified by homologous viral proteins that share evolutionary descent but have both lost inherited and gained new linear motifs (to be discussed in greater detail in the next section) [24882001]. In addition, not only can motifs convergently evolve in non-homologous proteins but combinations of motifs can also evolve in this manner, suggesting that (in contrast to domain architecture) functional necessity is more important than evolutionary descent [15585523].

1.4 Challenges in discovering linear motifs

Although expected to be abundant motifs are hard to discover. The low complexity that enhances their function and evolvability result in problems in discovering these motifs. The first SLIMs were discovered via carefully designed experimental gene fusion studies, e.g. ER retention signal or cyclin degron [1373379, 1846030]. Later, it became a common practice to look for new instances of known motifs. One can scan all proteome or proteins of interest to predict motifs for experimental follow-up. However, there are many dangers of using this very simplistic approach that were reviewed recently by Gibson et al [26581338]. The same amino-acid sequence may be functional depending on structural context. For example, a sequence matching nuclear export signal, that has 4 hydrophobic residues, can be often found in the hydrophobic core of globular domains. Experimental mutagenesis of such motif in a nuclear protein causes it aggregate which prevents its export from the nucleus which can erroneously be taken as evidence of a functional motif [23900254, 26581338]. This example highlights the importance of looking for motifs within disordered regions.

As demonstrated in a recent study by Hagai et al low complexity of motifs is a problem when predicting instances of known motifs proteome-wide [24882001]. Low complexity motifs can be found in similar frequencies in both true and randomised viral sequences leading to high numbers of false-positive motifs. For example, only 1-6% of known motifs matching protein in 2 viral species and 2 viral families occur in less than 0.1% of the randomised sequences.

1.5 Computational methods are necessary

Integration of multiple data sources, computational methods is essential for both finding novel instances of known motifs and discovering new motifs. Various ways of restricting the search space can solve the problems highlighted in the previous section. Accounting for structural context, residue conservation, known protein interactions all provide additional evidence for each particular motif

[26581338, 25555723]. Motifs that have contradicting evidence should never be subject to experimental follow-up: location in globular domain, absences of residues conservation even among related species. The last aspect is hard to get right because motifs can change location within a protein, poor quality of most protein sequence data and because alignment programs do not handle natively disordered sequences very well [26581338, 18460207]. Another problem for computational discovery of SLIMs are low complexity regions: long stretches of amino acid. However, masking them may introduces false negatives because these regions often serve as a substrate for SLIM evolution [28805808].

Computational de-novo discovery of motifs aims to address the challenge of low complexity of motifs as well as finding previously unknown motifs. Defining the search space is a key. Sets of proteins assumed to contain motifs can be derived from homologous sequences or protein interaction datasets [26581338, 25555723], both can be fairly noisy. Proteins interact with many other proteins via different regions in their sequence. So, rarely all known interactors contain the same motif. In addition, there is a problem of true but off-target motifs. Using any PPI network, we may discover true motifs using the wrong interactions with no recognition domain on the other side [21879107]. Finally, there might be non-functional motifs that are computationally discoverable and are recognised by correct domains in vitro but proteins never interact in-vivo. To address this problem, researcher may look whether proteins are ever expressed in the same cell type or cell type of their interest before doing experimental follow up.

While scientific community has generated more high-quality genomes, protein interaction data [28514442] and developed better tools for computational discovery [25792551] and in-vitro high throughput experiments [26297553], functional validation remains a challenge [26581338].

1.6 De-novo discovery of human linear motifs convergently evolved in viral proteins

Eukaryotic viruses rely on convergent evolution of SLIMs for interfacing and hijacking cellular function. This has been demonstrated in numerous targeted studies (discussed in section 3.7) and a recent systematic computational prediction of SLIMs in all viral proteins [24882001]. Interestingly, prokaryotic viruses do not use motifs as often because bacterial proteins generally have less motifs [24882001]. Not only do human-targeting viruses mimic cellular motif but the same study demonstrated that these motifs likely evolved ex-nihilo.

In our study, we take advantage of this property of viral proteins. We use experimental viral-human interaction data to define set of either human or viral proteins (Figure 3.4.1 B or C) that may contain motifs of interest. We then restrict motif search space by searching only for motifs convergently evolved in viral proteins. We follow the best practice of masking the disordered regions; however, we also estimate domains likely to mediate interaction to improve sensitivity and interpretability of motif prediction.

2 Methods

2.1 Protein interaction data

Protein-protein interaction (PPI) data was downloaded (FTP) from IntAct database release of 13 November 2017 [24234451] using loadIntActFTP function included into R programming language package MItools. Entries in IntAct database have to be cleaned of tags and textual description to facilitate further analysis using the cleanMITAB function. We use UniProt accessions to name participants of the interaction and filtered only protein-protein interactions. Custom taxonomy tree-aware code was used to identify which interactions in the database are human-human (taxonomy ID 9606, fullInteractome function), viral-viral (taxonomy ID 10239) and which interactions are between human and all viral taxa (taxonomy ID 10239, interSpeciesInteractome function). Taxonomy data was downloaded using Uniprot REST API (March 2018, loadTaxIDAllLower function). We keep isoforms and post-processed chains rather than defaulting to a canonical sequence. This may be especially important for some viruses whose proteins are all translated as a single polypeptide chain but then cleaved into functional proteins [26096987].

In addition to IntAct database, we used the BioPlex project data [28514442 and unpublished data] that includes around 7500 affinity purification experiments AP-MS experiments to identify more than 70000 interactions. The data was downloaded from the BioPlex website on 1 December 2017 (BioPlex 2.3) using the loadBioplex function. We used the mapping from Entrez gene ID to UniProt accessions that BioPlex provided. This mapping includes one gene to many proteins and many genes to one protein mappings. As a result, the interaction network has some interactions that were not actually tested. BioPlex may have a higher rate of "off-target" motifs (discussed in section 1.5).

We have used several subsets of data in the IntAct database: 2 large-scale studies and 2 interaction detection methods.

Large-scale studies. Data from two large-scale studies were selected using subsetMITABbyPMIDs function: Mann dataset [26496610] and Vidal dataset

[25416956, unassigned1304]. Mann study created 1,330 stable HeLa cell lines expressing 1,125 distinct bait proteins to be used for AP-MS. Vidal study was performed using yeast two-hybrid method.

Interaction detection method. Two subsets of IntAct database were based on interaction detection method: two-hybrid and affinity purification followed by mass-spectrometry [26681426, 26496610]. We define two-hybrid method using PSI-MI ontology: detection method = "transcriptional complementation assay" (MI:0018) - all methods which belong to this type (which are children terms in the ontology). We identify AP-MS method using two PSI-MI ontology terms: detection method = "affinity chromatography technology" (MI:0004) and participant identification method = "partial identification of protein sequence" (MI:0433). The use of ontology terms for searching interaction allows to specify only a single term as opposed to listing every individual detection method used. To identify which methods were included one can browse ontology lookup service [https://www.ebi.ac.uk/ols/ontologies/mi]. subsetMITABbyMethod function uses MI ontology to determine all child terms of categories described above to filter interaction data.

Randomised network. To obtain a control dataset for motif search we created a network identical in the number of edges and degree for each interacting protein but containing random interactions. To do that, we permuted interactions in the second position (IDs_interactor_B). We randomised BioPlex and viral-human network that were used in motif search downstream.

Commented [EP8]: These miught not only be Y2H but might be MAPPIT, KISS and others.

Commented [VK9R8]: That's true. I was aiming to include all these methods.

Commented [EP10]: It might be helpful to give a table that has the overview of the datasets with numbers etc. I assume you have to add these datasets either as an appendix or an external cd or sth, so you can include the name or the table legend in this table.

Commented [VK11R10]: I don't think I can add data on a CD or something. For your reference, this should be reproducible using scripts in the viral_project repo. All my datasets are also stored on EBI filesystem.

Table 2.1

Protein interaction network datasets

Dataset	N proteins	Unique interactions
Viral-human network	882 viral / 4544 human	14484

Human network: all IntAct data	19573	156732
Human network: BioPlex	12070	73665
Human network: Mann's data	4952	15601
Human network: Vidal's data	8638	44747
Human network: two-hybrid data	13552	69298
Human network: affinity-purification mass-spectrometry data	11707	59153

2.2 Degree distribution analysis

We analysed degree distributions of human and viral proteins in the human-viral, human-human and viral-viral protein interaction network. The degree is the number of interacting partners that a protein has. I loaded interaction datasets as described in section 2.1 using the loadHumanViralPPI function. I counted the number of interacting partners of viral proteins, human proteins or viral-targeted human proteins in each network or its subset by method or study (function humanViralDegree). I calculated the median of each distribution and visualised each distribution using the density plot (Figure 3.3.1 and Supplementary Figure 1). In addition, I calculated the number of interactions and the number of proteins in each network. The results of this analysis are discussed in section 3.1.

2.3 Protein sequences and domain prediction

2.3.1 Protein sequences

I used the Uniprot REST API (application programming interface) and Uniprot FTP [PMC5210571] to download sequences in FASTA format (20 October 2017). I used the downloadFastaMixed function [MItools package] to download sequences of all proteins, including protein isoforms and post-processed chains. This function downloads all canonical and isoform sequences in the SwissProt using UniProt FTP. Then, it loads the non- SwissProt sequences one-by-one using UniProt REST API

(downloadFasta function). Finally, a second function downloads post-processed chain region position in a protein sequence given post-processed chain IDs and subset protein sequence using the position of that region (downloadFastaPostproc function).

2.3.2 Domain prediction using InterProScan

For all human and viral proteins that have interaction data, available domains were identified/predicted using InterProScan software and InterPro sequence signatures. InterPro is a meta-database that collects sequence signatures of domains, families, sites and repeats [PMC5210578]. We run InterProScan version 5.25-64.0 in standalone mode on LSF computing cluster x86_64-pc-linux-gnu (64-bit, 8 cores, 16 GB RAM) running under Red Hat Enterprise Linux Server 7.3 (Maipo). We used the following versions of all databases: CDD-3.16 [27899674], Coils-2.2.1, Gene3D-4.1.0 [PMC5210570], Hamap-201701.18 [25348399], MobiDBLite-1.0, Pfam-31.0 [26673716], PIRSF-3.02 [19455212], PRINTS-42.0 [https://doi.org/10.1002/047001153X.g306301.pub2], ProDom-2006.1 [12230033], ProSitePatterns-20.132 and ProSiteProfiles-20.132 [23161676], SFLD-2 [24271399], SMART-7.1 [29040681], SUPERFAMILY-1.75 [11697912], TIGRFAM-15.0 [12520025]. The following code was used to start InterProScan:

```
/path_to/interproscan-5.25-64.0/interproscan.sh -i ./data_files/all_human_viral_proteins.fasta -f gff3 -iprlookup -goterms -b ./processed_data_files/all_human_viral_protein_domains102017
```

The output was saved in GFF3 format standard for storing sequence ranges.

2.3.3 Removing redundant domains and storing protein-domain pairs

Most InterPro member databases contain signatures of many types (domains, families, sites and repeats) so there was a need to filter domain signatures only. I downloaded which InterPro sequence signature is of which type from InterPro FTP using function getInterProEntryTypes. I parsed InterPro output file using function

readInterProGFF3, combined these files (addInterProEntryTypes function) and selected domains (SubsetByInterProEntryType function).

Several InterPro member databases may each contain signatures describing essentially the same sequence domain. We relied on InterPro work of integrating signatures from multiple databases to remove this redundancy. If two InterPro member databases provide a signature that matches the same domain, InterPro would indicate this by providing single InterPro identifier (e.g. IPR002048). We used this method and collapseByInterProID function to keep only one domain region per protein and InterPro ID. Note, this method keeps all domains that belong to the same family. For example, a generic protein kinase domain and tyrosine or serine-threonine protein kinase domains.

Next, we converted domain ranges in protein sequences and their annotations (Granges-class object in R) to a table of domain-protein pairs. We refer this as a domain-protein network.

2.3.4 Combining human domains with viral-human interaction data

In the next data-processing step, we merged domain-protein network and viral-human protein interaction network. Following that we calculated a number of descriptive statistics. These included how many proteins contain each domain (background count); the background frequency of a domain; how many human proteins are targeted by viral proteins; how many human interactors of a viral protein contain each domain (also referred to as domain count); what is the fold enrichment of a domain among interactors of a viral protein. Dependencies between these measures were examined (not shown but included into the .Rmd file). These measures used in estimating which domains are likely to mediate interaction.

2.4 Estimating which domains are likely to mediate interaction

Edwards et al have demonstrated that true motifs are often predicted using wrong interaction data, “off-target motifs” [21879107]. Experimentally determined

interactors of protein A yield a motif, however, this motif does not mediate the interaction of protein A with motif-containing proteins. Instead, protein B recognises this motif in a subset of interactors of protein A. In theory, estimating which domains are likely to mediate interaction should improve on-target prediction of SLIMs. We can evaluate motifs by examining their candidate recognition domains (section 3.7). In addition, we can filter motifs search datasets that have a strong recognition domain prediction (Figure 2.4).

We developed a method to identify domains enriched among human proteins targeted by a single viral protein. Enriched domains can serve as a proxy for domains likely to mediate interaction. These domains may be recognizing SLIMs in viral proteins, binding viral proteins via domain-domain interactions or show enrichment for functional reasons.

Rather than calculating a probability of observing a specific domain N times among the interactors of a viral protein given its background count, we calculate a probability of any domain being present in N times among the interactors of that viral protein (Figure 3.3, permutationPval function). We calculated this by sampling human targets of viral proteins keeping the degree and the total number of interactions intact; as if viral proteins were choosing human proteins irrespectively of their domain composition. For each viral protein, we compute how many times we see each domain. Then, we count how often permuted domain count is greater or equal to the observed domain count. This provides an empirical p-value – the probability of observing a domain count as high or higher under the null hypothesis. Permutation background is calculated for each viral protein to account for different number of interactions these protein form.

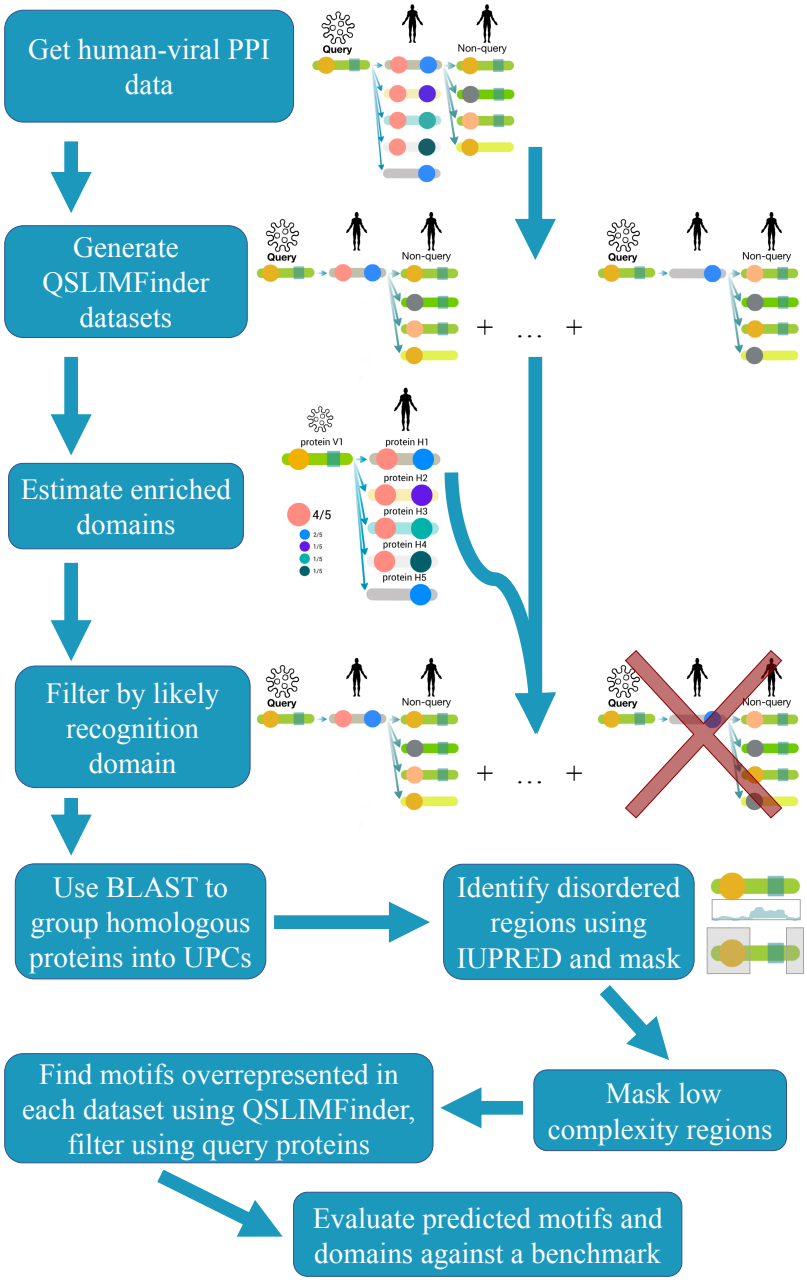


Figure 2.4. Flow diagram of the motif search workflow.

By not explicitly including the background frequency of a domain into calculation we improve the robustness of domain enrichment compared to count-based hypergeometric test. Edwards el at [21879107] discussed the problem of using hypergeometric test for finding enriched motifs: lack of compositional uniformity of the proteome, the difference in protein length. In our case, rare domains would be necessarily enriched in any set of viral-targeted human proteins even at a count of 1 or 2 because of the small number of interactors proteins usually have.

2.6 Motif search tools and setup

QSLIMFinder [25792551] is a command line tool that inputs: protein sequences as a FASTA text file, name of the query protein as a text file, other options not specific to each dataset. In our analysis, each dataset is defined by combination of a viral-targeted human protein (referred to as a seed) and a viral protein used as a query (Figure 2.4). We used the sequences of either viral proteins (Figure 3.4.1 B) or human and viral proteins (Figure 3.4.1 C) that interact with that single viral-targeted human protein. If that human protein has more than one viral interactor each of those interactors are used as a query. QSLIMFinder tool excludes the query sequence from the set of sequences used to calculate motif overrepresentation statistics. Motifs are represented as regular expressions. The probability of observing a number of regular expression matches by chance in a given set of protein sequences is calculated. Query protein sequence is used to filter the set of regular expressions being evaluated. This improves the sensitivity by reducing the number of multiple hypotheses being tested.

Table 2.6 summarises combinations of datasets and other options that we tested ranked based on performance in benchmarking.

Table 2.6

QSLIMFinder datasets that were tested

Dataset ID	Query network	Main network	cloudfix	Performance rank
qslimfinder. Full_IntAct3 cloudfixF.FALSE	Viral-human network	All IntAct data	FALSE	2
qslimfinder. BioPlex3 cloudfixF.FALSE	Viral-human network	BioPlex	FALSE	2
qslimfinder. all_viral_interaction3 cloudfixF.FALSE	Viral-human network	Viral-human network	FALSE	1
qslimfinder. randomised_BioPlex3 cloudfixF.FALSE	Viral-human network	Randomised Bioplex	FALSE	4
qslimfinder.randomised _all_viral_interaction3 cloudfixF.FALSE	Randomised viral-human network	Randomised viral-human network	FALSE	4
qslimfinder. Full_IntAct3.FALSE	Viral-human network	All IntAct data	TRUE	2
qslimfinder. Vidal3.FALSE	Viral-human network	Vidal data, published and not	TRUE	3
qslimfinder. all_viral_interaction3. FALSE	Viral-human network	Viral-human network	TRUE	1

2.6.1 Motif search software

We have used QSLIMFinder command line tool which is a part of SLIMSuite version released by Edwards group on 2016-09-12 [17912346, 25555723]. Homologous sequences are more likely to contain the same amino acid sequence patterns and therefore can artificially inflate support for each motif. QSLIMFinder groups homologous sequences using NCBI BLAST 2.6.0 [20003500] to produce Unrelated Protein Clusters (UPC or UP). In addition, short linear motifs are usually located in disordered regions, so disordered regions are masked using IUPRED protein disorder prediction software received on 4 September 2017 [15769473]. I compiled this software from source on LSF computing cluster x86_64-pc-linux-gnu running under Red Hat Enterprise Linux Server 7.3 (Maipo).

2.6.2 Creating motif search datasets

As described in section 2.6, Figure 2.4 and 3.4.1, motif search datasets are defined by the seed and query proteins. We had to create QSLIMFinder input two files for each dataset: a FASTA file containing sequences of proteins that interact with seed and a text file containing the identifier of a query protein. To create these files and generate BASH commands that would launch QSLIMFinder with these files and additional options we integrated protein interaction data, protein sequence data and domain data. This pipeline is implemented as the PPInetwork2SLIMFinder function.

We used as a seed all viral-targeted human proteins or proteins that have at least one domain predicted to be necessary for interaction with at least one viral protein at a p-value threshold of 0.5. This threshold was chosen empirically based on benchmarking. Choosing more stringent thresholds did not improve our recall as much as did removing proteins with no likely interaction-mediating domains (results section 3.5). Out of these seed proteins we chose those that had protein sequence data (section 2.3.1).

Filtered seed protein list was then used to produce datasets for QSLIMFinder as shown in Figure 2.4 (listInteractionSubsetFASTA function). When a seed human protein had more than one viral interaction I added other viral proteins to the non-

query set. Next, this list of QSLIMFinder datasets was filtered to include those where query protein has an enriched domain at a given threshold and set the minimal number of sequences in each dataset (1 viral query and 2 viral or 2 human non-query). Finally, files containing sequences and query protein names were created and paths to these files recorded.

2.6.3 Running interactome-wide motif search

To generate BASH commands calling QSLIMFinder on each dataset I combined the list of files paths with the path to QSLIMFinder software and other options (mQSLIMFinderCommand function). An example command looks like this:

```
bsub -n 1 -q research-rh7 -M 100 -R \"rusage[mem=100]\" python path_to/slimsuite/tools/qslimfinder.py blast+path=path_to/ncbi_blast_2.6.0/bin/ iupath=path_to/iupred/iupred dismask=T consmask=F cloudfix=F probcut=0.3 minwild=0 maxwild=2 slimlen=5 alphahelix=F maxseq=800 savespace=0 iuchdir=T extras=2 resdir=path_to/output/interactors_of.A0FGR8.P0DOE9./ resfile=path_to/output/interactors_of.A0FGR8.P0DOE9./main_result seqin=path_to/input/fasta/interactors_of.A0FGR8.P0DOE9.fas query=path_to/input/query/interactors_of.A0FGR8.P0DOE9.fas
```

A number options we used will be discussed in this paragraph (defaults used elsewhere). Disorder region masking was used (dismask=T) with the default 0.2 iupred score cutoff. Conservation masking was not used because motif search was done using non-homologous viral proteins. All motifs below QSLIMFinder Sig probability cut-off of 0.3 were retained (probcut=0.3). We used default motif length (the number of non-wildcard positions, slimlen=5) and the number of consecutive wildcards (minwild=0 maxwild=2). Longer motifs can be still discovered as a set of several shorter motifs. I limited the number of sequences in one dataset to 800 because large datasets can take a very long time to be analysed (maxseq=800). iuchdir=T tells QSLIMFinder to look for IUPRED disorder prediction software as provided by iupath argument rather than in an environmental variable. To tell

QSLIMFinder to generate occurrence file the following two options were used: savespace=0 and extras=2.

We tested both the option to restrict output to clouds with 1+ fixed motif (cloudfix=T) and not to restrict (cloudfix=F). Motif clouds are groups of motifs that overlap in 2 non-wildcard positions. Some clouds include only one ambiguous motif and Edwards recommends to remove these motifs [17912346]. When we added these motifs, we discovered more true motifs at more lenient thresholds. On the other hand, this approach adds more false-positive / candidate novel motifs making the precision and recall metrics very similar for both options.

2.7 Benchmarking instances of motif

2.7.1 Dataset

To evaluate if we are able to predict short linear motifs we tested how well we predict a set of known linear motifs in viral proteins. We gathered all linear motifs in viral proteins that were curated into Eukaryotic Linear Motif (ELM) database as of November 2017 [26615199]. This dataset contained regular expression that defines a motif and instances of 243 motifs in 143 viral proteins. Out of these we selected linear motifs in viral proteins that are known to interact with human proteins. We included ligand-binding, post-translationally modified and docking motifs while excluded degrons, cleavage and targeting motifs. These types of motifs tend to be more generic and present in many proteins. For example, targeting motifs can be present in viral and human proteins because of their shared location but not because they mediate interaction with the protein of interest - this makes them easy to discover but not relevant for our study.

The final benchmarking dataset contains 51 viral proteins. For every set of motif search datasets, benchmarking dataset is further trimmed to include only those proteins that we have searched for motifs. The largest benchmarking set we used contains 52 motifs in 35 viral proteins. This dataset is constructed using both viral-

human interactions and interactions of the viral-targeted proteins in the human network (Figure 3.4.1 C).

To benchmark our prediction of domains likely to mediate the viral-human interaction we used a list of known short linear motif binding domains annotated in ELM database. The major function of these domains is to mediate interactions. We would expect that a correct procedure for predicting domains likely to mediate interaction should predict SLIM-binding domains as likely to mediate interaction more often than other domains. 118 of these domains are present in 1016 human proteins targeted by 597 viral proteins.

2.7.2 Benchmarking pipeline

The goal of benchmarking was to determine which motif search options work best and to select a threshold at an acceptable precision and recall. To do that we found which motif instances discovered at a lenient QSLIMFinder Sig threshold of 0.3 match known instances from ELM database. Discovered unique motif instances (by range position) should match at least 2 amino acid positions of known motif instance. This could be further improved by evaluating non-wildcard positions in regular expressions.

First, I loaded enriched domain data and motif datasets prepared for QSLIMFinder. I optionally filtered both datasets by domain probability (results section 3.5-3.7). Next, I selected de-novo discovered motifs that were discovered using filtered QSLIMFinder datasets and ELM instances that could have been discovered using these datasets. ELM instances were filtered for specific motif types and redundancy in ELM motifs was removed. I did not merge two motifs if the motif type was different. As a next step, I calculated a joint predictor that includes by domain and motif p-values: (1 – motif_p_value) * (1 – domain_p_value). This predictor did not improve performance on known motifs (results not shown) suggesting a more sophisticated approach for integrating these is needed (discussed in section 3.8.2). We used motif p-value in all analyses.

I used motif p-value for each unique motif instance as a predictor of binary output: match to a known true motif vs a false positive or a candidate new motif. Several predicted motifs may correspond to one known motif such as 3 variants of PDZ motif in Figure 8, section 3.7.2.

I analysed the performance at different cut-offs using ROCR R package and a custom mBenchmarkMotifsROC function to organize my analysis. I examined precision, recall, true positive rate, the false positive rate at multiple cut-offs. I used this analysis to select three motif p-value cut-offs: lenient cut-off at 0.3; optimal cut-off at minimal p-value when precision is greater than recall (varies across datasets); and a stringent cutoff when precision is greater than 0.5 (varies across datasets).

I visualized results using Venn diagram plot (VennDiagram R package [citation("VennDiagram")]). These were hierarchically arranged into Figure 5 and Figure 6.

2.7.3 Examples of recovered and candidate motifs

We chose to examine recovered and candidate motifs predicted at a stringent cutoff using a combination of viral-human and human-human protein interaction networks (IntAct) and filtering by domain (Figure 3.4.1 C and 2.4). I evaluated whether domains likely to mediate interaction for each motif-containing viral protein are known SLIM-binding domains. To visualize results using Cytoscape, I have transformed data output of the benchmarking pipeline into a directed network: human protein –> recognition domain –> motif –> viral protein. I used motif p-value and domain p-value to scale the node size. Cytoscape file containing networks displayed in section 3.7: https://github.com/vitkl/viral_project/blob/master/results/thesis%20example%20plots.cys.

2.8 Motif pattern similarity

To compare the similarity of motif pattern for all motifs discovered at a stringent threshold using IntAct dataset (qslimfinder.Full_IntAct3.FALSE) we compared regular expression defining discovered motifs to all known motifs in ELM database. We used Comparimotif V3.13.0 software to do all pairwise comparisons and record motif similarity [18375965]. Motif complexity can be described using information content (IC). IC describes how much reduction in uncertainty is provided by a motif pattern. I run Comparimotif as command line tool included into SlimSuite (discussed in section 2.6.1) with default settings. The output of this tool was loaded in Cytoscape. I used the heuristic score of 1.162 (number of Matching Positions x Normalized IC) to filter motif similarity network [https://github.com/vitkl/viral_project/blob/master/qslimfinder.Full_IntAct3.FALSE/result/comparimotif.compare.cys].

2.9 Data analysis in R

Data analysis and data processing were performed using the language of statistical programming called R. Custom functions were written when needed and included in R package MItools [https://github.com/vitkl/MItools]. All analysis steps were performed and described using R Markdown reproducible documents where analysis code is complemented with textual description (*.Rmd). Any plots or other output produced by the analysis code as well as the details about R environment were included in the output document (*.html) when the analysis was performed. These analysis documents, some input data, output data and results of the project were organised in R package called viral_project [https://github.com/vitkl/viral_project]. The following files cover analysis described in previous chapters:

- interactions_and_sequences.Rmd: section 2.1, 2.3.1 and 2.3.2
- remove_redundant_domains.Rmd: section 2.3.3
- map_domains_to_human_viral_network.Rmd: section 2.3.4
- what_we_find_VS_ELM_count_justFisher.Rmd: section 2.4
- Motif_search_strategies_IntAct_Vidal_viral2.Rmd: section 2.6

- compr_benchmarking_strateg_IntAct_Vidal.Rmd, compr_benchmarking_strateg_cloudfixF_IntAct_BioPlex.Rmd: section 2.7 and 2.8
- compr_benchmarking_venn.Rmd: section 2.7.2, 2.7.3
- Degree_distribution_in_the_network.Rmd: section 2.2

3 Results & Discussion

3.1 Degree distribution in human and human-viral protein interaction network

3.1.1 Viral-human network is asymmetric: viral proteins interact with more human proteins than human proteins interact with viral proteins

To better understand interactions between viral and human proteins at a systems level we examined trends in the number of interactions these proteins to form. In addition, Figure 1 serves to summarise the protein interaction data we used in our study (number of proteins, number of interactions and how these are distributed across proteins). Both human-viral network and the human-human network are the largest to be used for motif discovery to date. We did not use any subsets of viral human data to retain as many interactions as we can although at some cost of quality.

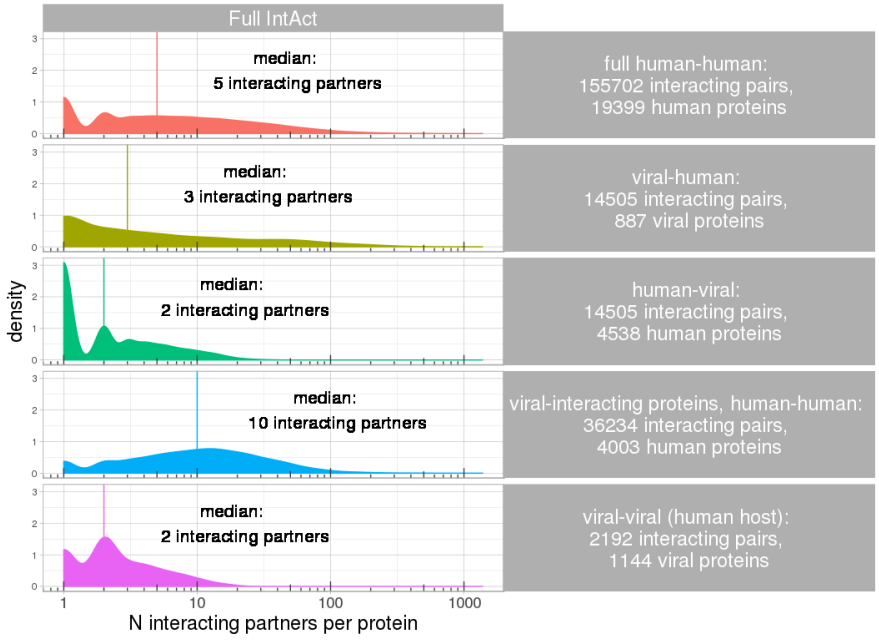


Figure 3.1.1. The distribution density of the number of interactions each human or viral protein has in each network that we used for our analysis. X axis shows the number of interacting partners for each protein, Y axis shows the distribution density. Different networks and different protein (viral or human) are shown in rows. Top row demonstrates that 19399 human proteins form a network with 155702 interactions with 5 interactions per protein on average (median). Row 2 and 3 show the distributions of the number of interactions in the human-viral network for viral and human proteins respectively.

We can observe four main trends:

- Both human and viral proteins in the human-viral network have on average fewer interactions annotated than human proteins do in the human network. Human-viral interactions are less studied. Less systematic studies were done to map out all interactions of viral proteins.

- Viral proteins interact with more human proteins that human proteins interact with viral proteins. This will be discussed later.

- Human proteins targeted by a virus have on average 2 times more interactions. To be discussed later.

- The human-viral data is very sparse: half of the human proteins have 2 or fewer interactions with viral proteins, half of the viral proteins have 3 or fewer interactions with human proteins. This suggests that many of these human-viral interactions alone will not provide enough information for motif discovery. This and a recent development of a specialised tool by Palopoli et al [25792551] motivated our approach of searching the human network for motifs using the sequences of each viral protein as a filter rather than searching for motifs in viral proteins only.

Let's discuss the second trend in more detail. We found that viral proteins tend to target many human proteins while human proteins are targeted by only a few viral proteins (Figure 3.1.1, row 2 and 3). This may reflect the biological need of viruses to interfere with multiple cellular processes. Alternatively, this difference may

reflect a technical aspect of studying viral interactions: more viral proteins may have been used as a bait because you need a viral infection or exogenous expression of viral proteins to detect these interactions. In addition, we see a general trend for viral proteins to have fewer interactions on average than for human proteins – which may be reflective of the same bias: much less large-scale human-viral studies were performed to date. 36 human studies used more than 50 baits while only a 5 viral did. This means bulk of the data comes from small scale targeted experiments rather than genome-wide protein interaction screens. Supplementary figure 2 shows that when testing human-viral interactions using two-hybrid interaction detection method viral proteins have many more interactions identified than human proteins, further supporting the hypothesis that lower numbers of viral interactions of human proteins can be partially explained by technical reasons.

3.1.2 Viruses target human proteins that appear as hubs only in the data biased for more well-studied proteins

The literature frequently tells that viral proteins target hubs (protein with many interactions) of the human network [PMC3593624, 25417202]. However, several reports have come out recently suggesting that fairly accepted association of disease relevance and a high number of interactions in the protein interaction network may be overestimated if researchers account for study bias [26300911]. The study bias means that better-studied proteins have more interactions defeating the usefulness of the protein degree as a measure of protein function. To address this problem and improve the coverage of the protein interaction networks several systematic studies profiled interactions between up to ~17000 proteins [26496610, 28514442, 25416956]. We can use the protein degree in each these studies as a better measure of the true functional importance of proteins.

Commented [EP12]: To cut down space this can also be described in 3 sentences

In the previous section, we saw that human proteins targeted by viruses tend to have much more interacting partners (median 10 compared to median 5 for proteins not targeted by viruses). This may be explained by functional difference and hub-

targeting behaviour of viruses or by technical and study bias. Affinity purification methods tend to extract and measure protein complexes rather than direct and binary interactions causing proteins to appear to have more interactions when measured using these methods. Many of viral-human interactions in our dataset come from this type of studies which may explain a part higher interaction numbers. The other possibility is that viral- targeted human proteins are more studied overall.

In Figure 3.1.2, We see that while affinity-purification followed by mass-spectrometry (AP-MS) was used to generate more interactions for viral-targeted proteins what two-hybrid - interaction detection method doesn't explain higher numbers of interactions entirely. Study bias does explain: both Mann and Vidal datasets have identical medians and identical shape of the distribution for virus-targeted and all human proteins. This tendency of viral proteins to interact with the most studied human proteins may be especially strong because many of these most studied proteins (including P53 [218111]) were first discovered due to their viral interactions (a major method of human protein discovery before the first human genome was sequenced).

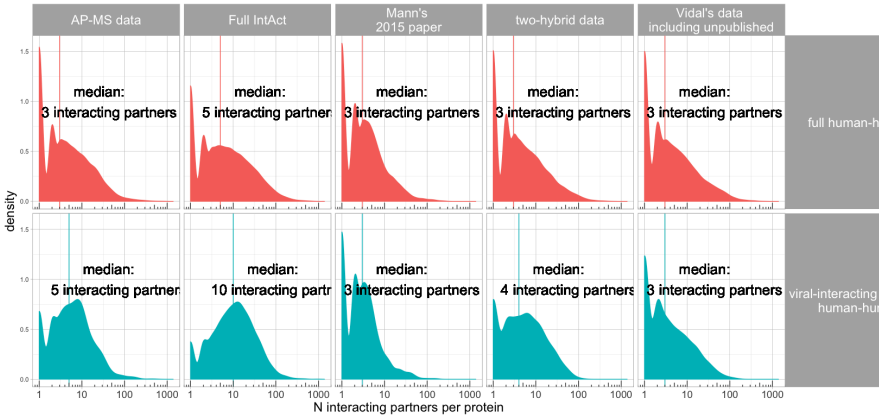


Figure 3.1.2. The distribution density of the number of interactions for all human proteins or viral-targeted human proteins shown for all data, or selected protein interaction detection methods and large-scale unbiased studies. X axis is

log10 scale and shows the number of interacting partners per protein; Y axis shows the distribution density.

3.3 Domains likely to mediate interaction are enriched in SLIM-binding domains

Knowing that viral protein targets human proteins containing a specific domain can allow filtering the data to increase the signal-to-noise ratio in the motif search downstream. To estimate that, we found domains enriched in human proteins that interact with each of viral protein [https://github.com/vitkl/viral_project/blob/master/processed_data_files/domain_res_count_20171019.RData].

Although it is common to predict domain-domain interactions using the protein interaction data [28514442], up to the best of my knowledge no attempts to predict domain-protein interactions were published. Hypergeometric distribution is commonly used to calculate the probability of overlap in elements between categories. In this case, the overlap between interacting partners of a viral protein and proteins with a specific domain under the null hypothesis. However, it provides a p-value for the enrichment of a specific domain but not for the enrichment of any domain, because this method depends on comparing the frequency of proteins with a specific domain in a set of those that interact with a specific viral protein to the background frequency. It has been shown that approaches that rely on the background frequency, such as all proteome, are usually bad at estimating the background distribution for identifying overrepresented motifs due to non-uniform sequence composition of the proteome [25555723, 25207816].

I suggest that hypergeometric distribution is also poor at predicting interaction domains. We saw that a low number of interactions can artificially drive up the frequency in a set (not shown). When a protein interacts with only 3 other proteins the minimal frequency of any domain will be 0.33 causing even the most abundant domains to be enriched in that set. For example, P-loop containing nucleoside triphosphate hydrolase is the most abundant domain in a background set of viral-targeted human proteins but its frequency is just 0.01; which means that the domain will be 5-fold enriched even if it is present only in 1 out of 20 proteins. These

problems make hypergeometric distribution not appropriate for identifying enriched domains.

To combat these problems, we developed a permutation-based procedure to calculate the probability of seeing any domain N number of times among the interactors of the viral protein (Figure 3.3 A). Unlike Fisher test, our procedure highlights the domains enriched in known SLIM recognition domains (one-sided Two-sample Kolmogorov-Smirnov test, D^- = 0.13548, p-value < 2.2e-16). Known SLIM recognition domains have preferentially low p-values (Figure 3.3 B). This allows us to use domain enrichment as a proxy for domain likely to mediate interaction (including motif-mediated interaction).

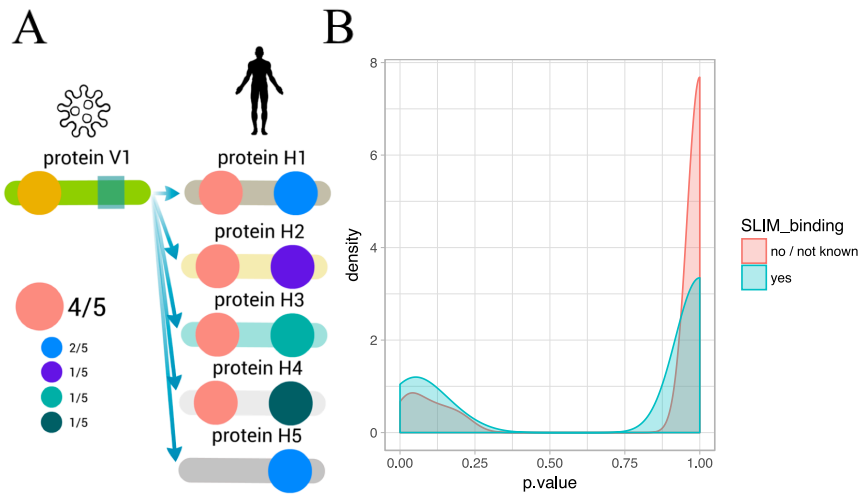


Figure 3.3. Identifying enriched domains. A. Diagram illustrating how enriched domains are identified. We find domains in human proteins that each viral protein targets. We count how many times each domain is observed. Next, we use a permutation-based approach to calculate the empirical p-value for any domain appearing that many times in proteins targeted by V1. B. Distribution density plot of empirical p-values for domains known to bind SLIM or for all other domains. The x-axis shows the p-values; Y-axis shows the distribution density. SLIM-binding domains tend to have more of low p-values than all other domains.

We see 2 main limitations of this approach:

1. Viral proteins may target functionally related human proteins. These proteins may have a shared domain architecture so we may not be able to distinguish which of them is more likely to mediate interaction. One example is a candidate SH3 domain-binding motif that binds 4 kinases with identical domain architecture (discussed later).

   Another example is P-loop containing nucleoside triphosphate hydrolase domain. If enriched it would reflect the preference of a virus to bind protein with GTP-ase activity; however, a different set of domains may be responsible for binding.
2. Some human-viral interactions are mediated by domain-domain interactions. Many of the enriched domains will be responsible for binding but will not aid the discovery of SLIMs.

We selected a p-value cut-off of 0.5 to exclude all domains that are not likely to mediate interactions with viral proteins. We used all other domain-protein pairs when constructing the motif search datasets: we only look for motif in a viral protein that has a likely recognition domain in the human protein. After filtering we were left with 5379 interactions between 396 viral proteins and 754 enriched human domains.

3.4 De-novo discovery of Short Linear Motifs

We identified SLIMs convergently evolved in viral proteins using a probabilistic method developed by Edwards et al (QSLIMFinder) [PMC4495300]. These motifs are over-represented in proteins that interact with viral-targeted human proteins. For this analysis, we assumed that each viral protein has a motif recognised by a globular recognition domain in a human protein. The same domain can recognise instances of this motif in human proteins (Figure 3.4.1 C) or in other viral proteins (Figure 3.4.1 B). The goal of both approaches is to discover the sequence of this motif.

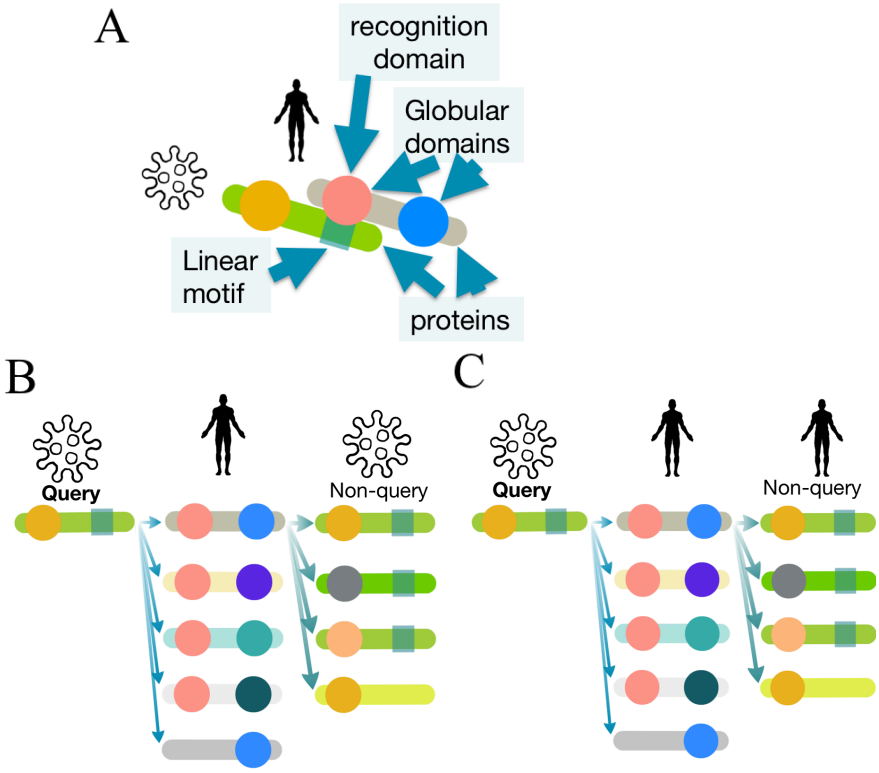


Figure 3.4.1. Schematic illustration showing how datasets for motif search were constructed. Non-query proteins were used to search for motifs that must be

present in query proteins. Each dataset consists of all interacting partners of one human protein and one query protein. A. Legend. B. Datasets can be constructed using human proteins that interact with multiple viral proteins. C. Query protein can be a viral protein that mimics a motif present in non-query human proteins. Adding these non-query proteins may provide more power and interpretability over the viral only dataset.

Rather than relying on FDR-adjusted p-value provided by QSLIMFinder as a measure of false discovery rate we evaluated the performance of our approach by comparing predicted motifs to known motifs at 3 different cut-offs (Figure 3.4.2). Known motifs were taken from ELM database as described in section 3.2. We evaluate performance at predicting motifs in viral query proteins, however, we also predict motifs in the human network.

We can recover known motifs using both strategies shown in Figure 3.4.1 B and C. Figure 3.4.2 gives the breakdown of the number of candidate motifs we discovered and known motifs we recovered. Although we applied the same criteria to the set the cut-off, the approach using only viral-human data requires lower adjusted p-value to recover the same fraction of known motifs at the same error rate than the approach including all human data from IntAct database [24234451]. We also used the human protein interaction data from large unbiased screen done by Vidal's group [25416956] (Supplementary figure 3). Vidal's data performs worse than all data from the IntAct database. This agrees with the previous study by Edwards that demonstrated that their motif search method (SLIMFinder) is more sensitive to the absence of signal than to the presence of noise [21879107, 20055997]. This means it is more important to keep as many proteins with a motif as possible even at the expense of adding more proteins lacking motif. It is better to have 10/100 than 3/6 proteins containing a motif.

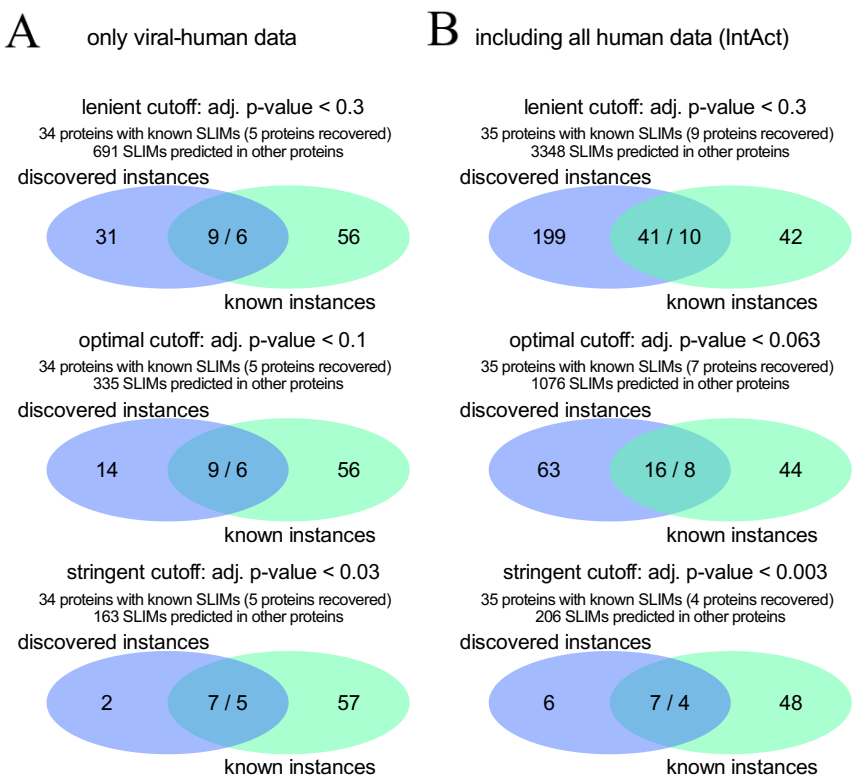


Figure 3.4.2. The number of candidate motifs we discovered and known motifs we recovered. Two datasets construction strategies (A and B) and 3 p-value cut-offs are shown. The blue circle shows the number of motif instances predicted but not known. The green circle shows the number of motifs instance known but not recovered. Overlap shows the number of discovered instances that match known instances (predicted/known). These numbers differ because several similar predicted motif patterns can match one known motif at the same location in a protein sequence (see section 3.7.2 for example). Adj. p-value is the QSLIMFinder Sig which is the probability of observing N motifs in a random sequence corrected for the number of all possible motifs tested. Lenient cut-off reflects the probability of false discovery as high as we are maximally willing to tolerate. At optimal cut-off, precision, the fraction of discovered instances that match known, is approximately equal to recall,

the fraction of known instances we recovered. At a stringent cut-off, precision is 0.5 or above, we discover on average either one candidate new motif or one false positive motif for every known motif in a viral protein. For each dataset and cut-off, we show how many motifs were discovered in proteins that do not contain known motifs.

Using the approach that included all human data we recovered 1/5 of known motifs present in query proteins that we could have found. However, we also discovered many motifs that don't match known motif: on average 5 candidate new motifs or 5 false positive motifs per each known motif in each protein. This number of new motifs per protein is rather unlikely: viral proteins contain at most known 4 motif instances. For example, genome polyprotein of Hepatitis C virus (P27958) has 4 instances of N-glycosylation motif (MOD_N-GLC_1) [26615199]. Viral proteins contain at most 3 known motifs of different classes (Early E1A protein of Human adenovirus C, P03255) [26615199]. For this reason, we considered two more stringent thresholds. We could trade off the lower power to discover true motifs for higher precision. At an optimal cutoff we missed 2 additional known motifs but reduce the number of potential false-positives; however, we still predicted as much as 4 candidate new motifs or 4 false positive motifs per each known. Finally, we chose a stringent cut-off under which we recovered only known 4 instances in viral proteins but had a lower candidate new/false positive rate. We will examine these motifs in detail in the section 3.7.1 and 3.7.2.

To illustrate how benchmarking based on known instances is helpful for selecting cut-off, let's look at false discovery rate-adjusted p-values at the most stringent cut-off. For the viral only approach, the stringent cut-off is QSLIMFinder Sig p-value < 0.03 which matches < 0.3 after FDR adjustment for testing multiple datasets. FDR-adjusted Sig p-value for the approach that includes human data is also above traditional p-value < 0.05 (0.078). This suggests that statistical model-based FDR may not reflect FDR on the real data. Also, different protein interaction datasets return true motifs with different p-values but still on top of the list.

At a stringent cut-off, both approaches discover overlapping but not identical sets of motifs. Out of combined 10 motifs, we recovered 7 motifs using viral dataset (match 5 known) and 7 motifs by adding human interaction data (match 4 known). By using the dataset that included human protein interaction network missed we missed a known retinoblastoma protein-binding motif (LIG_Rb_LxCxE_1) in Protein E7 of Human papillomavirus and Early E1A protein of Human adenovirus C [http://elm.eu.org/instances/LIG_Rb_LxCxE_1/P03129/21, http://elm.eu.org/instances/LIG_Rb_LxCxE_1/P03255/118, 26615199]. By using viral only dataset, we missed known nuclear localisation signal motif in Polymerase basic protein 2 of Influenza A virus and a fragment of the known PDZ-domain-binding motif in Protein E6 of Human papillomavirus [http://elm.eu.org/instances/TRG_NLS_Bipartite_1/P03428/738, http://elm.eu.org/instances/LIG_PDZ_Class_1/P06463/153, 26615199]. This suggests that while in some cases human network provided a signal in other cases it added noise.

Finally, we compared the human network produced by a single unbiased study by Vidal's group [25416956, unpublished] to that of full IntAct. At equivalent stringency thresholds, we can discover slightly fewer motifs and the same precision threshold requires lower p-value (Supplementary figure 3). This suggests that Vidal's data may be depleted of SLIM-mediated interaction compared to all human interaction data. Yeast two-hybrid screens take two proteins out of cellular context that may be required for motif binding (e.g. phosphorylation). In addition, Vidal's group demonstrated that their method recovers interactions that are on average stronger (higher affinity of binding) than those recovered by other methods, such as affinity-purification mass-spectrometry [unpublished].

As a next step, we are combining motifs returned by viral only and all human data approaches and using a less noisy human PPI network, such as BioPlex [28514442], that may capture motif-mediated interaction better than Vidal's two-hybrid screens. In section 3.6 and 3.7 I am going to focus on the results obtained using full IntAct network discussed in this section.

As illustrated in this section, even at a stringent threshold we can recover know motifs and predict as much as 206 motifs in viral protein that do not contain known motifs. Next, we would like to know if filtering the interaction data by the presence of a likely interaction-mediating domain can improve our power to recover motifs.

3.5 Filtering by domain improves sensitivity of motif prediction

Many of the known viral instances of motif do not have enough support in the protein interaction data to be rediscovered. These motifs are not recovered even at a lenient cut-off (Figure 3.4.2). For many other motifs, we do not have enough information in the interaction data to point to likely recognition domain. Among those that have enough - we can discover a higher fraction of true motifs (Figure 3.5). Although we captured one less of the true positive motifs at a stringent threshold using both approaches, we recovered a higher fraction of known motifs compared to the approach without filtering for domains.

This demonstrates that the approach is working and even at a stringent threshold we can recover true motifs. In the next sections 3.6 and 3.7, we will discuss recovered motifs and a number of candidate motifs we identified using an approach that includes human interaction data. First, we will look at motif similarity, then, examine specific instances in detail.

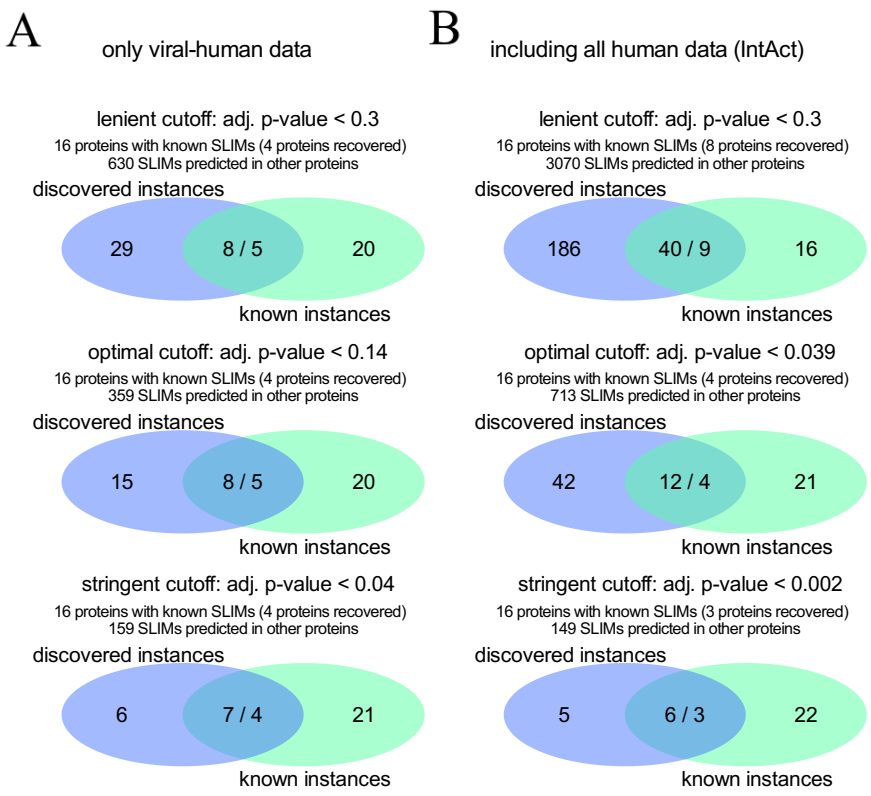


Figure 3.5. The number of candidate motifs we discovered and known motifs recovered when filtering by domain. Two datasets construction strategies (A and B) and 3 p-value cut-offs are shown. The blue circle shows the number of motif instances predicted but not known. The green circle shows the number of motifs instance known but not recovered. Overlap shows the number of discovered instances that match known instances (predicted/known). These numbers differ because several similar predicted motif patterns can match one known motif at the same location in a protein sequence (see section 3.7.2 for example). Adj. p-value is the QSLIMFinder Sig which is the probability of observing N motifs in a random sequence corrected for the number of all possible motifs tested. Lenient cut-off reflects the probability of false discovery as high as we are maximally willing to tolerate. At optimal cut-off, precision, the fraction of discovered instances that match

known, is approximately equal to recall, the fraction of known instances we recovered. At a stringent cut-off, precision is 0.5 or above, we discover on average either one candidate new motif or one false positive motif for every known motif in a viral protein. For each dataset and cut-off, we show how many motifs were discovered in proteins that do not contain known motifs.

3.6 De-novo discovered motifs are similar to known motifs

To explore which candidate short linear motifs we discovered, we examined which known motifs are similar motifs selected under the stringent threshold. All of these motifs resemble known motif in the ELM database (matching sequence with 0.5 information content). We summarise similarity results further filtered above score 1.162 to avoid excessive assignments in Figure 3.6.1.

We see a clear cluster of nuclear localisation signal (targeting) motifs that have KR-containing pattern. We also see proline-rich motifs (P..P.[HKR] and P..P.P.D) recognised as SH3 domain ligands.

Pattern similarity comparison is very prone to overprediction and at the same time may not capture low information content motifs (very few defined positions). For example, classic C-terminal PDZ domain-binding motif was not recognised.

Whether other pattern similarity matches of candidate motifs are likely to represent the correct motif class requires further investigation. A more robust way of identifying motif classes into match motif pattern similarity and whether these motifs have a correct matching recognition domain.

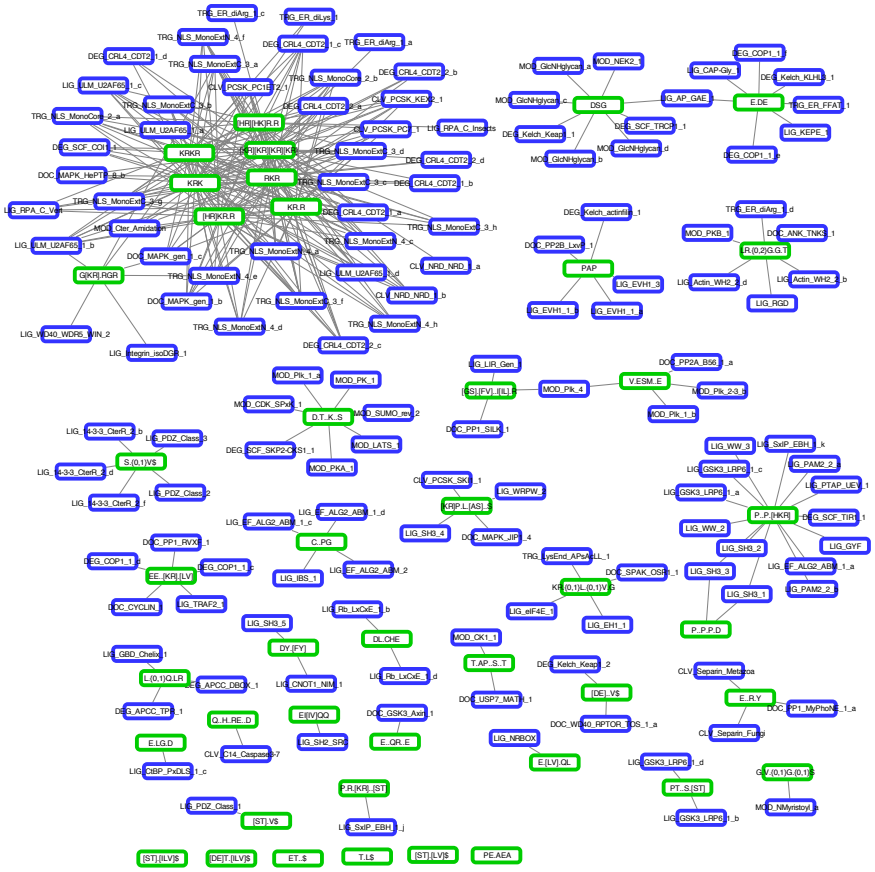


Figure 3.6.1. The network of similarity of the pattern of discovered short linear motifs to known motifs. Green nodes are discovered motif patterns, blue nodes are known SLIMs in ELM database, edges show the similarity between motifs above the threshold of 1.162 (Score, Comparimotif3).

3.7 Examples of recovered and candidate motifs

3.7.1 Several classes of candidate motifs recovered and predicted alongside and their likely recognition domains

To examine hits produced by a combination of human-viral and full human-human (IntAct) networks I selected candidate motifs under the most stringent threshold (precision > 0.5 or 1 candidate motif per each known motif). This dataset is more interpretable because in addition to a motif that has convergently evolved in a viral protein it predicts motifs in the human protein used to bind the same recognition domain. I manually examined predicted motifs and their most likely recognition domains. These results are summarised in Supplementary Table 2.

Encouragingly, the most prevalent group of candidate motifs were not targeting motifs but classic C-terminal ligand-binding motifs ([ST].[LV]$) recognised by a PDZ domain. 26 variants of these motifs ([ST].[LV]$, [DE]T.[ILV]$, ET..$, [ST].V$, [DE]..V$, T.L$, [ST].[ILV]$) were predicted at C-terminus of 16 viral proteins that bind to 7 PDZ-domain-containing human proteins. 2 instances of these motifs were already known and will be discussed in the next section. For all of these instances except (3 in 2 proteins), PDZ-domain was correctly identified as the most likely or one of the most likely domains mediating the interaction.

As expected, one of the most prevalent groups of candidate motifs were 12 nuclear localisation signal (targeting, TRG, KR-rich motif) motifs present in 11 viral proteins that bind to 4 human proteins (nuclear import machinery) each containing an Armadillo-like domain - correctly picked up by our domain enrichment procedure. Many of other viral proteins used in our study are localised to the nucleus but were not found to interact with nuclear import machinery suggesting that the motifs observed may be mediating a more stable interaction. An alternative explanation is that viral proteins containing motifs are abundant enough to bind enough copies of nuclear import machinery for these interactions to be detected, a hypothesis supported by the fact that 5 of these motifs are located in capsid proteins. Domain enrichment also picked up the Armadillo-like domain as the most likely for

a couple of motifs that do not resemble nuclear localisation signal (E..QR..E, D.T..K..S, PT..S.[ST], V.ESM..E). We found two of these motifs using interactions of human proteins that do not belong to nuclear import machinery (26S proteasome non-ATPase regulatory subunit 1 - E..QR..E motif; E3 ubiquitin-protein ligase HUWE1 - PT..S.[ST] motif). Further investigation of these candidate motifs is required to determine if the Armadillo-like domain of these proteins may be binding non-canonical ligands.

Instances of several other ligand-binding motif classes were predicted: 4 candidate WD40 motifs, 1 SH3 motif, 1 EF-hand motif, 1 PH domain-like. Some of these motifs are further examined in the following sections. Some of these candidate motifs have domains that are not known SLIM-binding domains as tagged as the most likely, including 1 Cyclin-like, 1 Keratin type 2 head, 2 Gro-EL-like and 7 BAG domain binding candidate motifs (refer to Supplementary Table 2). However, both motifs and domain may have been found in error, so further investigation is needed prior to experimental validation.

Next, we will discuss how individual motif instances are supported by our analysis and independent literature.

3.7.2 Successfully recovered 2 known PDZ-domain binding motifs

At a stringent confidence threshold (precision > 0.5) we have recovered 2 known PDZ-domain-binding motifs in Protein E6 of Human papillomavirus (HPV) type 16 and 18 (Figure 3.7.2 A and B respectively).

In this example, the same protein in two related viruses targets an overlapping set of human proteins. Protein scribble homolog (SCRIB) and tyrosine-protein phosphatase non-receptor type 3 (PTPN3) are targeted by both viruses.

ELM annotation of this motif is based on structural evidence: the interaction of Protein E6 motif with PDZ domain of the MAGI1 human protein. Although we searched for motifs in proteins that bind MAGI1 both before and after filtering for domain – we could not recover any even at a lenient threshold. Nonetheless, this

ELM instance of a PDZ-domain-binding motif is recovered using 3 other PDZ-domain proteins. Let's examine how HPV can hijack these.

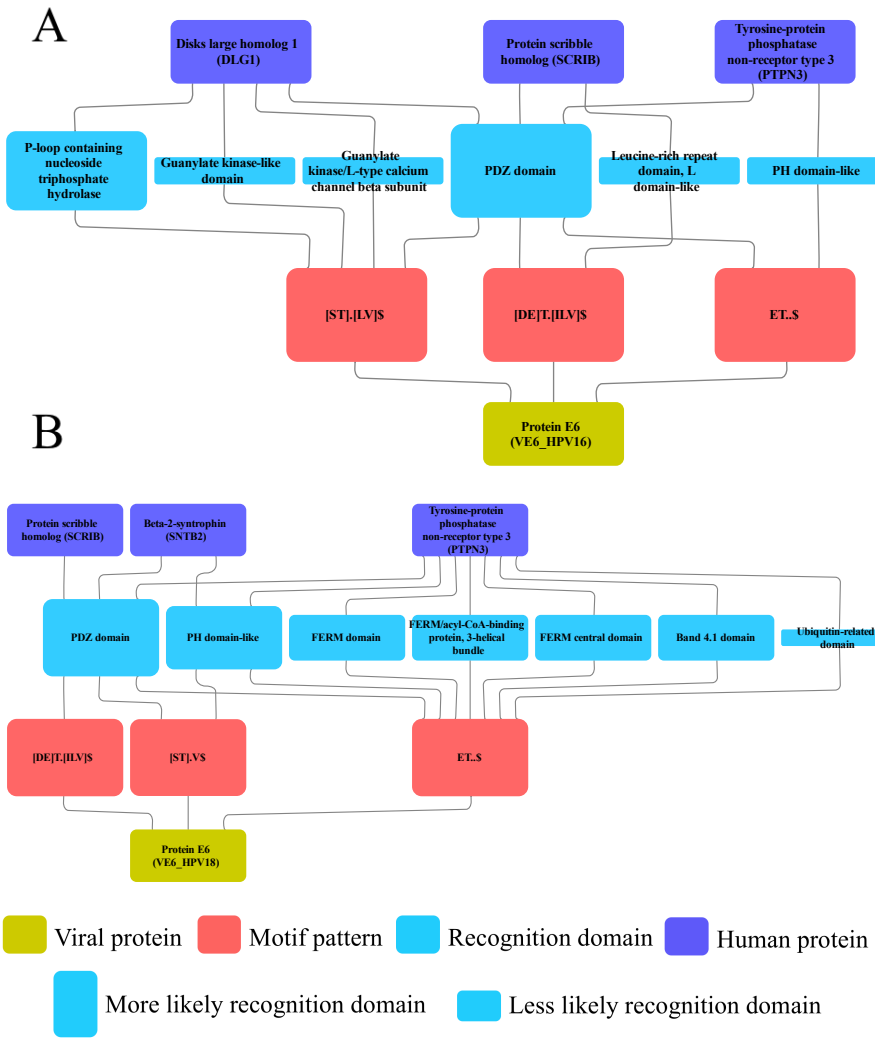


Figure 3.7.2. Known PDZ motifs in protein E6 of Human papillomavirus 16 and 18. We show motif variants and which domains in which human proteins may be responsible for binding protein E6. Three variants of this motif were predicted in

protein E6 and in human proteins that interact with viral-targeted human proteins called DLG1, SNTB2, SCRIB and PTPN3. The PDZ domain is the most enriched among targets of protein E6 and is the domain that mediates interaction with this known motif. A. The network of this motif in Human papillomavirus 16. B. The network of this motif in Human papillomavirus 18.

E6 protein of HPV bind a range of PDZ domain-containing human proteins. PDZ-binding C-terminal region varies in different types of these viruses which dictates binding preference [PMC4970744]. HPV 16 and HPV 18 E6 target the highest number of human proteins. Some human proteins such as DLG1 are targeted by all types of HPV E6 proteins. Targets of E6 proteins, including DLG1, PTPN3 and SCRIB, are usually ubiquitinated and degraded by the proteasome [PMC2748042, 10523825, PMC1865939]. E6 proteins accomplish this by acting as scaffold proteins to recruit E3 ubiquitin ligase E6-AP (UBE3A) via LXXLL motif located in the ligase [17023019, PMC1865939]. A major disease relevant function of E6 protein lies in activating telomerase to immortalise infected cells. This function also relies on the ubiquitin-targeting role of E6 protein to degrade a repressor of telomerase transcription NFX1-91 [15371341]. Other studies also suggest that E6 has a more complex relationship with its targets: SCRIB protein appears to positively transcription and translation rates of the E6 protein [29074188]. At this point, it is also not entirely clear how targeting proteins involved in establishing apical-basal cell polarity and intracellular contacts is beneficial to this virus. Either way, interactions of HPV E6 with human PDZ-domain proteins are crucial for infection and tumorigenesis.

3.7.3 PDZ-domain binding candidate motifs

As mentioned in section 3.7.1, PDZ domain motifs are the most abundant in our set of discovered motifs. We have recovered 2 instances annotated in ELM and 14 instances that are not. Here we will examine candidate motifs that have the most support.

The first motif, like all the known motifs, is located E6 protein but in a different HPV type: HPV-70 (Figure 3.7.3.1). Although not annotated in ELM this motif is also known. According to the study by Thomas et al C-terminal peptides of HPV-70 E6 bind fewer proteins than that of HPV-16 or HPV-18 that were discussed earlier [PMC4970744]. According to their results HPV-70 E6 motif indeed binds to DLG1, however, it does not bind SCRIB and none of the motifs tested bind ERBIN [PMC4970744, Figure 2]. This may suggest that the interaction of the HPV-70 E6 protein with SCRIB is indirect or spurious. We can still identify the PDZ domain motif in this protein using interactions of SCRIB because SCRIB indeed binds PDZ domain motif; however, the PDZ domain of SCRIB may be selective enough to avoid binding HPV-70. The absence of peptide interaction with ERBIN is inconclusive because none of the peptides tested in that study bind ERBIN, which may be indicative of ERBIN not being expressed in the cell line (HaCat) where researchers performed the peptide pull-down assay. Overall, this study by Thomas et al serves as a validation of this PDZ motif instance in HPV-70 E6 protein but it points to a potential discrepancy in the protein interaction data used for our study.

In terms of the function, ERBIN serves as an adaptor protein that binds unphosphorylated ERBB2 receptor thus stabilising this state [16203728]. It is essential for localising ERBB2 to the basolateral side of epithelial cells [10878805]. Given that HPV also targets other proteins related to apical-basal cell polarity, such as DLG1 and SCRIB were discussed earlier, ERBIN may represent a feasible target.

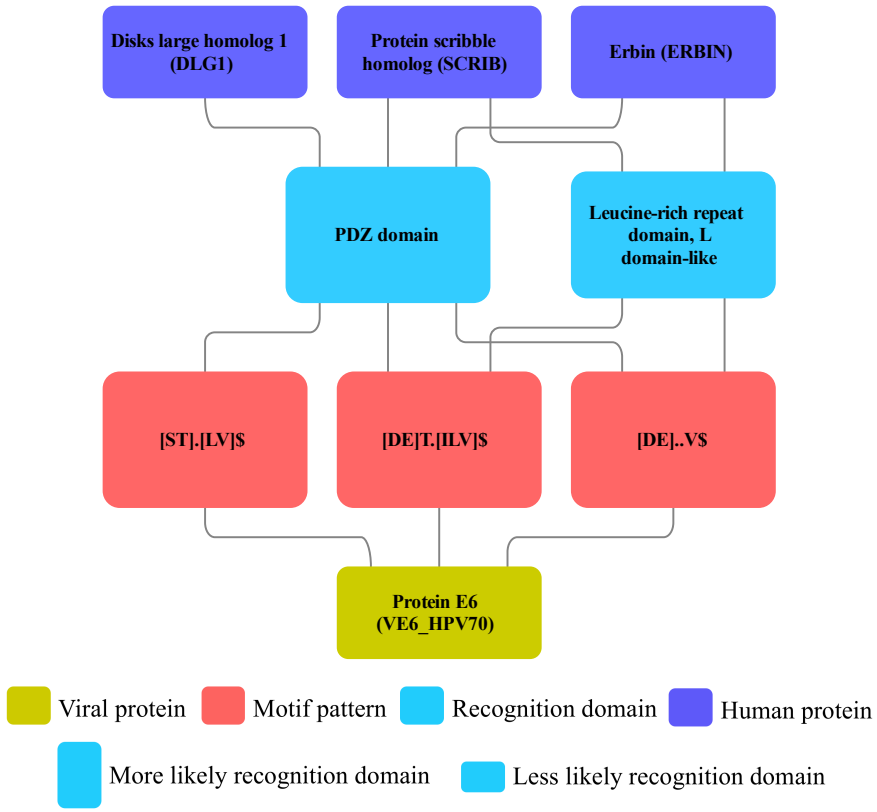


Figure 3.7.3.1. Candidate PDZ motifs in protein E6 of Human papillomavirus 70 was validated in a previous study but not annotated in ELM. We show motif variants and which domains in which human proteins may be responsible for binding protein E6. Three variants of this motif were predicted in protein E6 and in 74 human proteins that interact with viral-targeted human proteins called DLG1, SCRIB and ERBIN. The PDZ domain is the most enriched among targets of protein E6.

The second candidate PDZ domain-binding motif that we predict is located in the non-structural protein of Influenza A virus H5N1 (Figure 3.7.3.2). This motif is discovered using datasets of 4 human proteins: SCRIB and ERBIN, DLG4 and GIPC1. Although not annotated in ELM, this motif in also known. Moreover, it was

demonstrated that this motif enables H5N1 to disrupt tight junctions via its interaction with SCRIB and DLG1 [21849460]. Motif patterns that we identify match overall pattern of highly pathogenic avian (RS.V) but not human (ES.V) Influenza A viruses [21247458]. H5N1 virus hijacks proapoptotic function of SCRIB using PDZ domain motif to change its subcellular localisation [PMC2953166]. This prevents the apoptotic death of infected cells. Interaction of NS1 with ERBIN, DLG4 and GIPC1 are not described in detail to date.

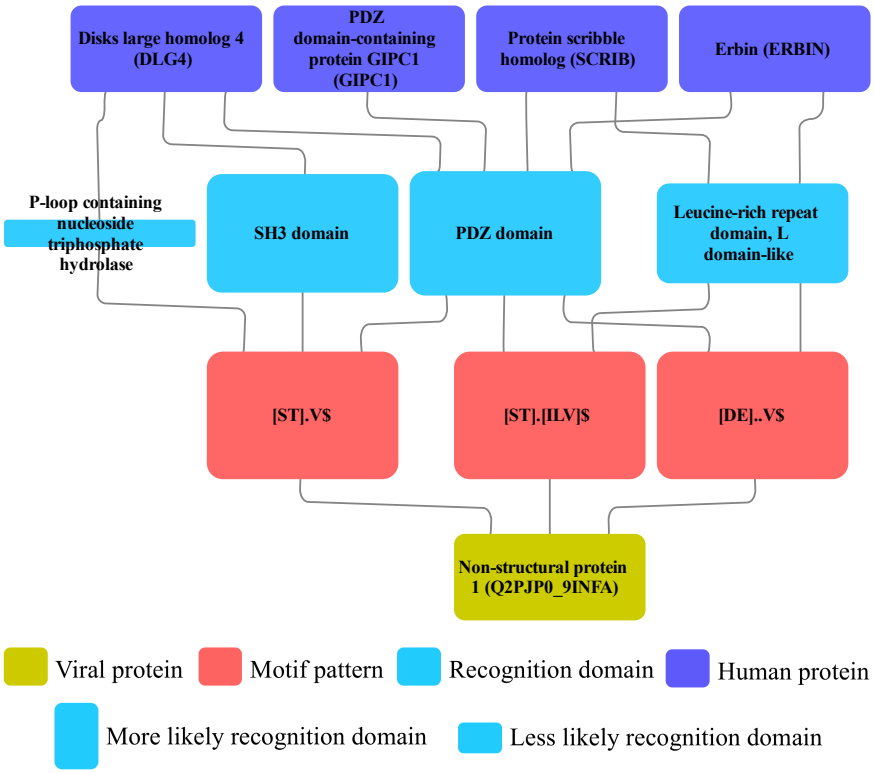


Figure 3.7.3.2. Candidate PDZ motifs in Non-structural protein 1 (NS1) of Influenza A H5N1 virus was validated in a previous study but not annotated in ELM. We show motif variants and which domains in which human proteins may be responsible for binding NS1. Three variants of this motif were predicted in protein NS1 and in 86 human proteins that interact with viral-targeted human proteins called

DLG4, GIPC1, SCRIB and ERBIN. The PDZ domain is the most enriched among targets of protein E6. The high enrichment of SH3 domain and Leucine-rich repeat domain may reflect functional preference of NS1.

To sum up, two PDZ motifs not annotated in ELM that had the most support in our analysis were already described in previous studies. This serves as a validation of our de-novo discovery pipeline. Next, let's examine a number of less abundant motifs.

3.7.4 SH3-domain binding candidate motif

We de-novo discovered an instance of SH3 domain-binding motif in Nef protein of Human immunodeficiency virus type 1 (Figure 3.7.4). Although we could not identify a single likely interaction domain we saw that P..P containing sequence resembles canonical SH3 domain ligand [7953536]. Nef is known to interact with only 5 human proteins 4 of which share a domain architecture of SRC kinase. This motif is indeed another known instance not annotated in ELM database. As confirmed by mutagenesis studies, P..P.[HKR] motif allows Nef to bind SH3 domain of a subset of SRC kinases to activate them and promote viral pathogenicity [PMC398106, 16849330].

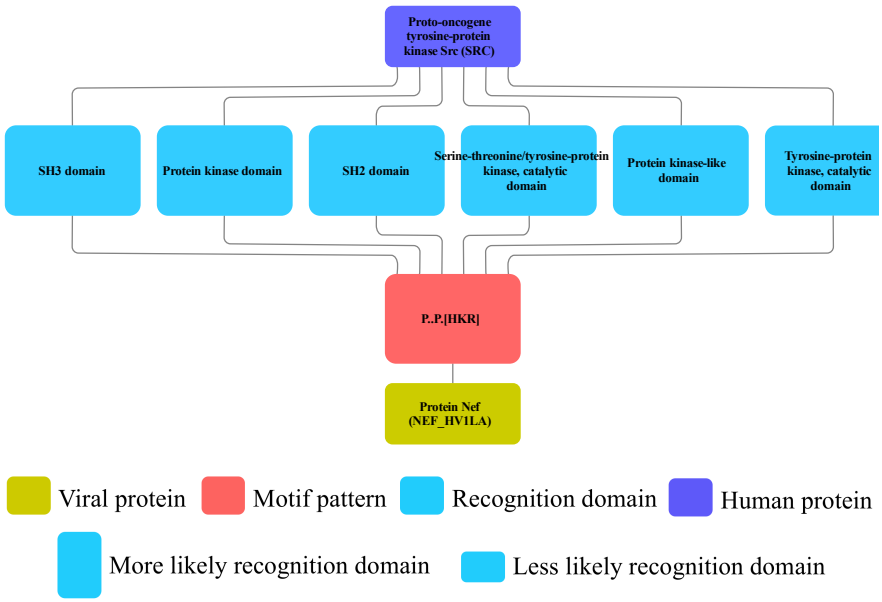


Figure 3.7.4. Candidate SH3 motifs in Nef protein of Human immunodeficiency virus type 1 was validated in a previous study but not annotated in ELM. We show motif variants and which domains in which human proteins may be responsible for binding Nef. One variant of this motif was predicted in protein

Nef and in 93 human proteins that interact with SRC. All domains except Tyrosine-protein kinase catalytic domain are equally enriched which may reflect the functional preference of Nef or bias in currently available data because Nef is known to bind only 5 human proteins.

SH3 domain-binding motif in Nef protein is yet another validated motif that was not included in our training data.

3.7.5 WD40-domain binding candidate motifs

We predict DSG motif as a candidate WD40 binding motif (Figure 3.7.5.1) located in four viral proteins of three viral species: Vpu protein of Human immunodeficiency virus (VPU_HV1H2 and VPU_HV1S1), Large T antigen of SV40 virus (LT_SV40) and non-structural protein of Rotavirus A (NSP1_ROTP5). This motif is recognised by two F-box/WD repeat-containing proteins: FBXW11, also known as β-TRCP1, and BTRC, also known as β-TRCP2. Both serve as a substrate recognition subunit of SCF (SKP1-CUL1-F-box protein) E3 ubiquitin-protein ligase complex. This complex ubiquitinates and targets proteins for proteasomal degradation [10648623, 10066435]. SCF (FBXW11 or BTRC complex) is a part of signalling pathways including Wnt-beta-catenin and NF-kappaB pathways where it targets either beta-catenin (phosphorylated by GSK3beta) or IkappaB for degradation. In turn, this inhibits (beta-catenin) or activates (NF-kappaB) the transcription factor at the end of the pathway [10321728, 10437795].

Vpu has a known DSG..S motif instance (DSGNES) that allows HIV to hijack SCF ubiquitin ligase to degrade host proteins such as the Antiviral Factor Tetherin/BST-2 and CD4 [PMC2729927, 9660940]. This motif (when phosphorylated at both serines) is normally recognised by FBXW11 and BTRC targeting motif-containing protein for degradation. Apparently, Vpu has found a way to escape self-degradation [9660940]. NSP1 of Rotavirus A also has a known DSG_S motif [28851847]. This protein uses host ubiquitin ligases to degrade key

host factors activating interferon production such as IRF3, IRF5 or IRF7 [27009959, 17251580]. Interferon is normally produced as a response to viral infection and helps to limit infection of neighbouring cells [24751921]. These cases serve as a validation of our motif discovery procedure: a true motif not represented in ELM database – the data we used to select optimal parameters and the threshold.

Large T antigen (TAg) of the SV40 virus is not known to have a DSG motif. This protein interacts with a tumour suppressor and a DNA damage sensor P53 (this is how P53 was discovered) [PMC353757]. A regulator of P53 a ubiquitin ligase MDM2 contains a known DSG motif and is itself degraded by β-TRCP1/2 discussed earlier [PMC3494375]. Although, TAg of the SV40 virus doesn't have a DSG..N motif validated (note the substitution of the last serine to asparagine) its homolog TAg protein of JC virus does contain DSG..S motif [PMC3017642].

To get a better idea of the structural aspect of this interaction I have performed docking of three peptides to β-TRCP1/ FBXW11 using pepsite 2 (PDB 1P22, chain A). This analysis indicates that a short DSG motif that we predict may have a binding site in FBXW11, however, not with very high confidence. In addition, DSG motif is predicted to bind F-box rather than WD-40 motif that we predict using domain enrichment procedure. Surprisingly, docking a full sequence of the known motif (DSGNES) of Vpu or TAg of SV40 motif (DSGHET) do not have a high-confidence binding site predicted by Pepsite 2 either.

Both DSG motifs have the very strong support of 24/36 (FBXW11) or 29/56 (BTRC) proteins with motif among non-redundant proteins that interact with FBXW11 or BTRC.

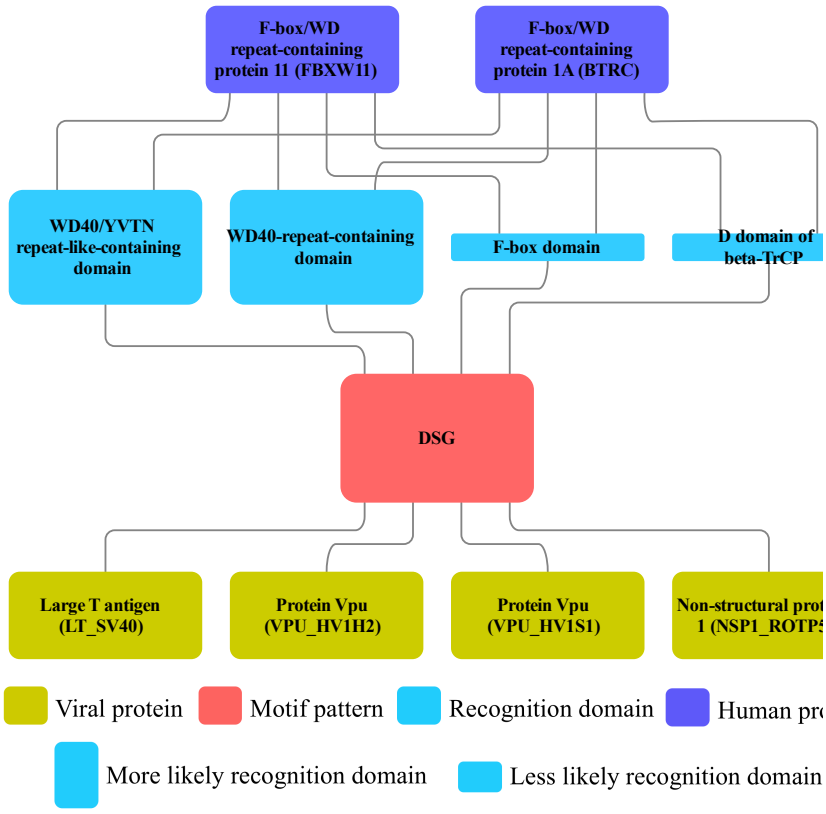


Figure 3.7.5.1. The network of candidate DSG motif. This motif is predicted in 4 viral proteins. All of them target 2 human substrate recognition subunits of SCF E3 ubiquitin ligase complex. 3 motif instances were validated in a previous study but not annotated in ELM. The exception is LT of SV40. WD40 domain is the most enriched among targets of protein VPU_HV1H2 and SV40.

The second candidate WD40-binding motif (Figure 3.7.5.2) was predicted in Polymerase basic protein 2 (PB2) RNA-polymerase of 2 Influenza A virus strains (B4URF7 protein in strain A/WS/1933 H1N1, C5E527 protein in A/New York/1682/2009 H1N1). We also predict this motif in 4 human proteins that all bind human Elongator complex protein 1 (ELP1). ELP1 is involved in elongation of

transcription by RNA polymerase 2 as a part of a complex that plays a role in chromatin remodelling and acetylates H3 histones [22854966]. WD40 prediction as the most likely domain is supported by 9/210 proteins or 5/140 proteins containing that domain (for each viral strain respectively). Given the RNA-polymerase function of PB2, we can hypothesize that it also hijacks host elongation factors via E.V..G.{0,2}N.{0,1}Q motif to facilitate this process.

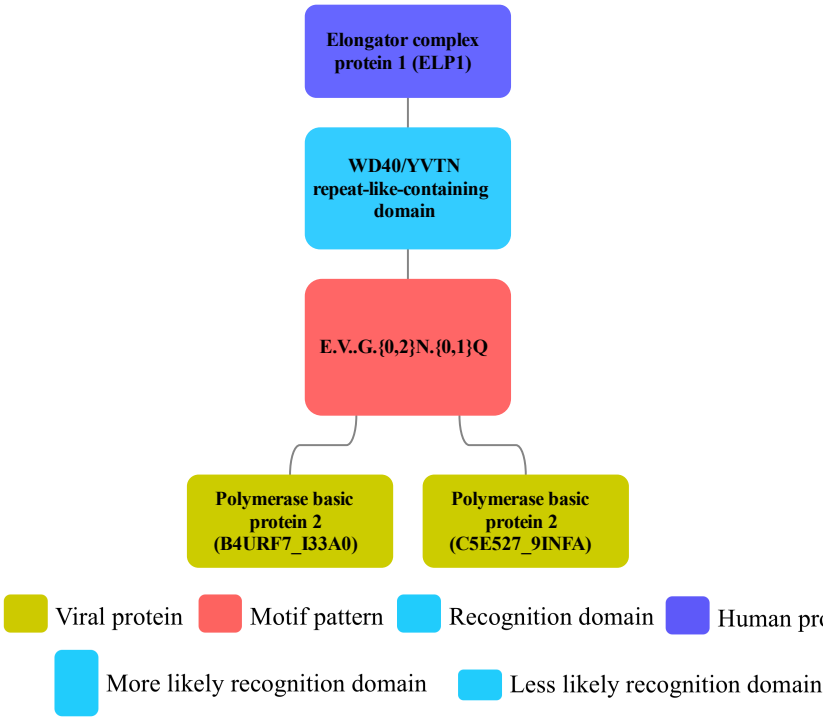


Figure 3.7.5.2. The network of candidate E.V..G.{0,2}N.{0,1}Q motif predicted in polymerase basic protein 2 in 2 Influenza A strains. We predict this motif to be recognised by WD40 domain in human protein ELP1.

3.7.6 Double-stranded RNA-binding domain and EF hand domain - candidate motifs

We predict LR.{0,2}G.G.T motif that may be recognised by double-stranded RNA-binding domain of Q96SI9 – human spermatid perinuclear protein that recognises viral RNA. We predict this motif in 6 non-structural viral proteins of 4 influenza A strains and 4 human proteins (Figure 3.7.6.1). These viral proteins are involved in blocking the translation of host mRNA [8525619] and also inhibit TRIM25 mediated ubiquitination which is a part of antiviral response [23209422]. We can speculate that this motif mimics RNA that this human domain normally recognises.

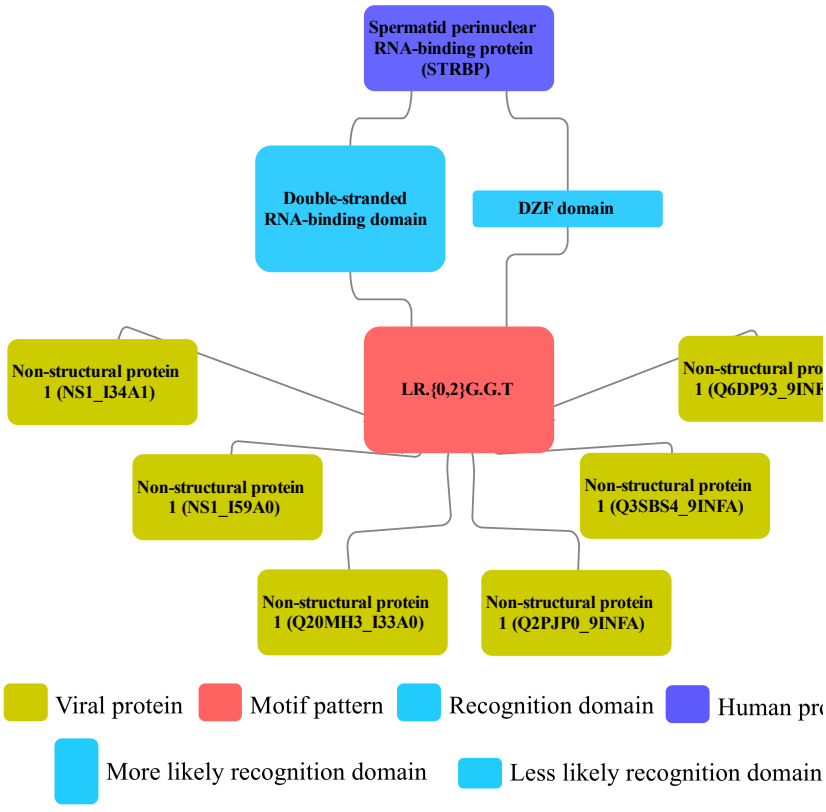


Figure 3.7.6.1. The network of candidate LR.{0,2}G.G.T motif predicted in 6 non-structural proteins of 6 Influenza A strains, including both avian and human

lineage. We predict this domain to be recognised by Double-stranded RNA-binding domain of the STRBP human protein.

EF-hand domain-binding motif is shown in Figure 3.7.6.2. Cab45 is an EF-hand domain and Ca(2+) binding protein which is necessary for the sorting of secretory proteins at trans-Golgi network. Cab45 oligomers bind secretory and plasma membrane proteins [27138253] and target them to plasma membrane / extracellular space. This protein is not known to interact with viruses besides a recent high-throughput effort to profile interactomes of multiple Influenza A virus strains [28169297]. Cab45 is targeted by 12 distinct viral proteins from 6 viral taxa, although, these interactions were profiled using affinity purification methods that measure both direct and non-direct interactions. A viral protein (PB1, Q5EP37) that targets Cab45 and contains EI[IV]QQ motif is one of the RNA-dependent RNA polymerases of Influenza virus and is an essential component of viral transcription machinery [23600869]. RNA polymerase proteins (including PB1) stays bound to viral RNA and are packaged into viral particles. We can hypothesise that Cab45 serves to facilitate this process. Although feasible, the confidence in this hit decreases because unlike DSG motif described earlier, this candidate EF-hand binding motif was discovered only in 4 out of 33 protein sequences (partners of Cab45) and the domain relevance is supported by only 2 out of 45 PB1-binding proteins (Figure 13). Examining 3 human proteins predicted to have this motif may clarify how likely this motif is to represent a real hit.

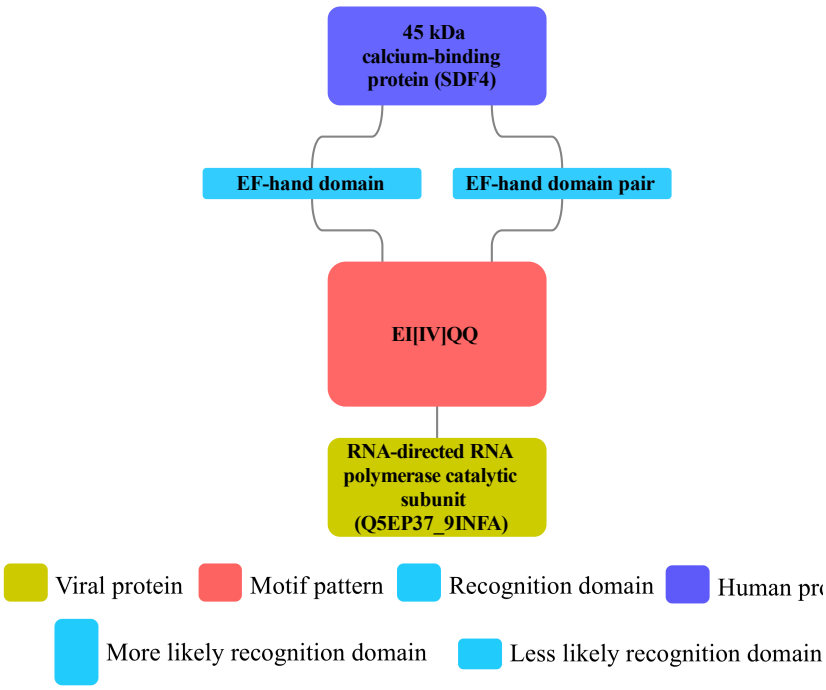


Figure 3.7.6.2. The betwork of candidate EI[IV]QQ motif located in Influenza A RNA polymerase and potentially recognised by the EF hand domain of the human protein SDF4.

3.7.7 BAG-domain binding candidate motif

We found a candidate motif (L.{0,1}Q.LR) potentially recognised by BAG domain in seven repeats in the Epstein-Barr nuclear antigen leader protein 5 (Figure 3.7.7). This motif is found in 13 other proteins that bind to human co-chaperone BAG2 and is also predicted as that mediating interaction with a related protein BAG3 but at a lenient significance level. The Epstein-Barr nuclear antigen leader protein 5 (EBNA5) is one of the first proteins detected during EBV infection and is essential for transforming B-cells by acting as a transcriptional co-activator [29462212, 16177824]. BAG2 and BAG3 are HSP70 and HSC70 co-chaperone proteins and work as nucleotide exchange factor [9873016]. This way EBNA5 may

enhance HSP70 and HSC70 chaperone activity or affect cell proliferation or apoptosis via functions of BAG2 or BAG3. Examining 13 other proteins in which this motif was predicted motif will provide more evidence if BAG domain may indeed recognise L.{0,1}Q.LR motif. However, a prediction of viral peptide (LGQLLR) binding on a PDB structure of mouse BAG3 (1uk5) or human BAG1 (3fzf) domain using pepsite 2 [22600738] does not indicate a presence of a strong binding site in this domain (http://pepsite2.russelllab.org/match?molvis=jsmol&pdb=1uk5&chain=A&ligand=LGQLLR).

Commented [EP13]: It is more useful and pretty to put the actual structure there, screenshot it or sth. The output of these will be deleted within a week after you ran it.

Commented [VK14R13]: This link will make pepsite2 run again rather than open the of my old job.

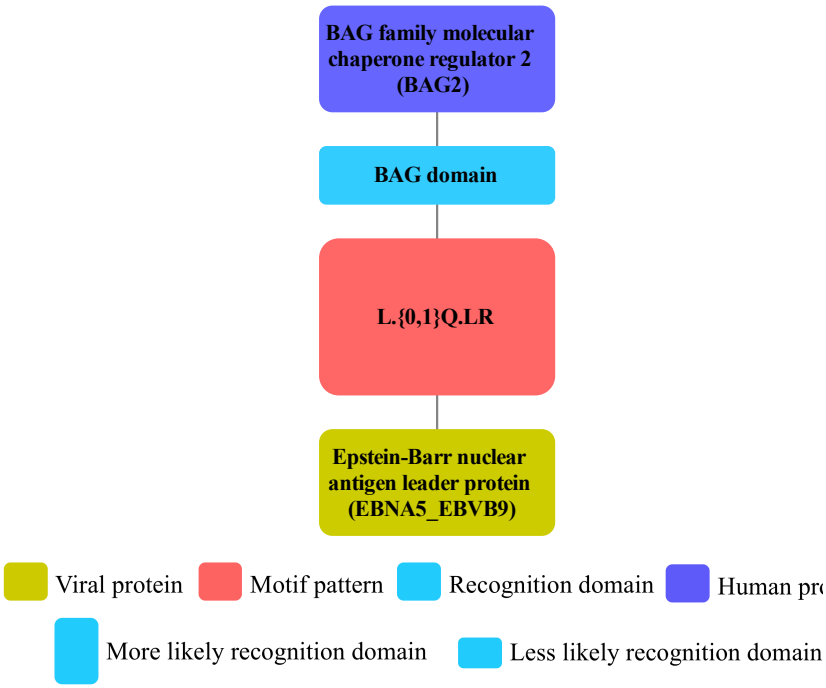


Figure 3.7.7. The network of candidate L.{0,1}Q.LR motif in the Epstein-Barr virus nuclear antigen leader protein. It is potentially recognised by the BAG domain of human HSP70 and HSC70 co-chaperone proteins.

3.8 Future directions

We showed that we can use protein interaction data and the property of viral proteins to convergently evolve host SLIMs to recover 3 known motif instances from ELM database and 6 motifs validated by previous studies. We also find candidate motif instances that are not known and predict likely recognition domains. In some cases, discovered motif resembles motif known to bind the most likely domain (PDZ-domain, SH3 domain, DSG motif), but in many cases, it does not. More work can be done to improve the precision of the domain prediction and the precision and sensitivity of motif predictions. In this section, we will discuss the approaches that we might use.

3.8.1 Motif-domain molecular docking and improved analysis on the human side

To improve both the motif and domain prediction, we can use motif-domain molecular docking using Pepsite 2 as we did in several post-hoc examples (DSG, BAG-domain motif) [22600738]. This can allow prioritizing motifs that have a good structural match to the domain, but also provide an independent way of evaluating most likely recognition domain. Two main limitations of this approach are the availability of domain structures and the low sensitivity of the method. For example, MAGI-1 PDZ domain was co-crystallized with HPV-16 Protein E6 [21238461], however, pepsite2 did not predict a strong PDZ-motif binding site on the surface of this domain (http://pepsite2.russelllab.org/match?molvis=jsmol&pdb=2KPL&chain=A&ligand=RRETQL). Computational speed is not a limitation, Pepsite 2 is fast enough to enable interactome wide docking (at least not slower than QSLIMFinder) to score as many motif-domain pairs as possible.

We can do more analysis on the human protein interaction network to improve likely interaction domain prediction which is currently done based solely on a human-viral network. Essentially, we predict domains for viral proteins only,

however, we identify motifs in human proteins too. We can improve domain prediction by considering both viral and human proteins that share the same motif. If 4 out of 5 proteins with a motif have the same domain enriched – this domain is more likely to mediate interaction. If all 5 disagree this could mean the motif itself is not functional.

In our analysis, we have not used protein sequence conservation to limit regions of proteins where we look for motifs. This conservation is generally recommended [], however, viral proteins evolve quickly so conservation filter may remove true motifs []. Nonetheless, we can use a conservation filter on the human proteins: a motif should be present in human and several other animals with well-annotated genomes. A potential problem may be that targeting of a motif by a virus applies selection on human motifs that may increase the evolutionary turnover rate of these motifs rendering a conservation filter ineffective, however, this has never been demonstrated.

These 3 suggestions can improve individual steps of our motif discovery pipeline. Next, we can integrate the predictions from each step in a smarter way.

3.8.2 Integrating multiple predictors: random forest

We can take a machine learning approach to integrating probabilities of a motif, a domain and their interaction predicted by pepsite 2. Each of these probabilities provides useful information about the motif we want to discover. By combining these we may improve both the sensitivity and the specificity of motif prediction.

The simplest integration approach is to assume independence of our prediction multiply p-values provided by each of the methods. Linear models provide a similar solution: the weighted sum of p-values. Both models have a disadvantage that a strong motif signal will be downweighed by a weak domain or a weak pepsite prediction. Given that these weak predictions can be driven by the absence of data rather than true biology we may be limiting our ability to recover known motifs. To combat that, we can use a decision tree-based method such as random forest or XXX

that is robust to missing values. These methods can learn both the strong signal from one source and combine 3 weak signals.

3.8.3 Experimental validation

The final step is the experimental validation of novel motif-domain interactions. As was discussed in the literature review, different classes of motifs would require different functional validation experiments. However, first, we need to validate a physical interaction. In this assay, interacting protein fragments are identified through NGS sequencing of phage genomes allowing high-throughput identification of domain-linear motif interactions. The main limitation is that each domain has to be in-vitro synthesised [28002650]. Our computational prediction highlights domain instances in human proteins that are worth screening against disordered regions of viral proteins.

Conclusions

1. I retrieved and processed experimental interaction data from public databases and the literature. I examined the properties of viral-human interaction network. Viral-targeted human proteins appear as hubs but this effect can be attributed to study bias in the aggregate protein-protein interaction datasets.

2. By using viral-human network, probabilistic motif search tools and the sequences of viral proteins to limit the search space we can recover known instances of short linear motifs in viral proteins and predict new candidate motif.

3. We identified protein sequence domains in all viral and human proteins. We estimated which human domains are likely to mediate interaction with each viral protein. These domains are enriched in domains known to recognise motifs. Filtering probable domains improves recall. Integrating domain and motif prediction improves interpretability of the results.

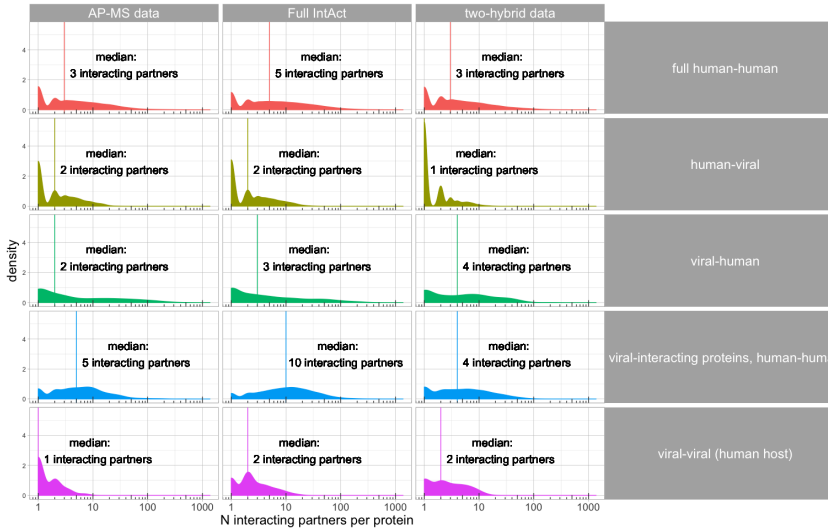
4. At a stringent threshold of 50% precision we can recover 3 known motif instances from our training set de-novo discover 6 known motif instances that were not in our training set. We predict 43 candidate novel motif instances. These motifs and their likely recognition domains will be experimentally validated using phage display.

5. I implemented this motif search pipeline in R statistical programming language, using command-line tools and LSF high-performance computing cluster. This pipeline can be used by the group (and scientific community) to predict motifs as new protein interaction data is generated.

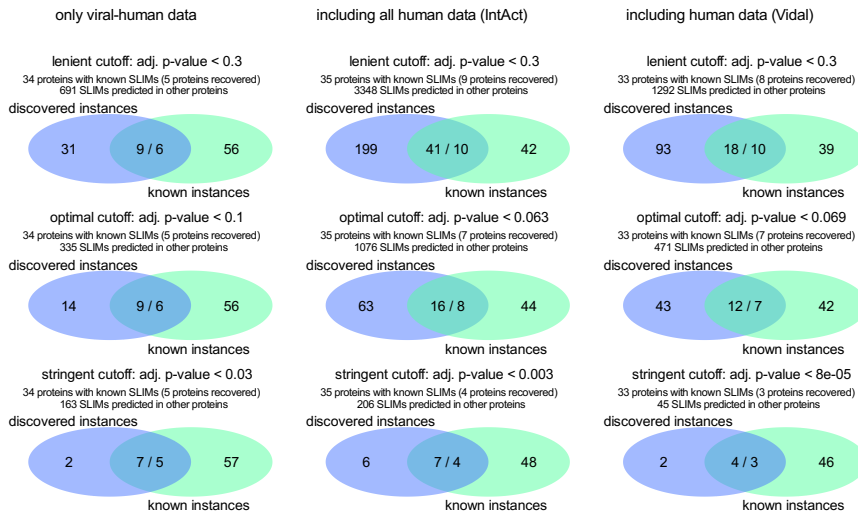
6. This work contributes to our understanding of the linear motif – recognition domain code and guides the choice of viral-targeted human domains for further experimental studies of disordered viral proteome.

Supplementary materials

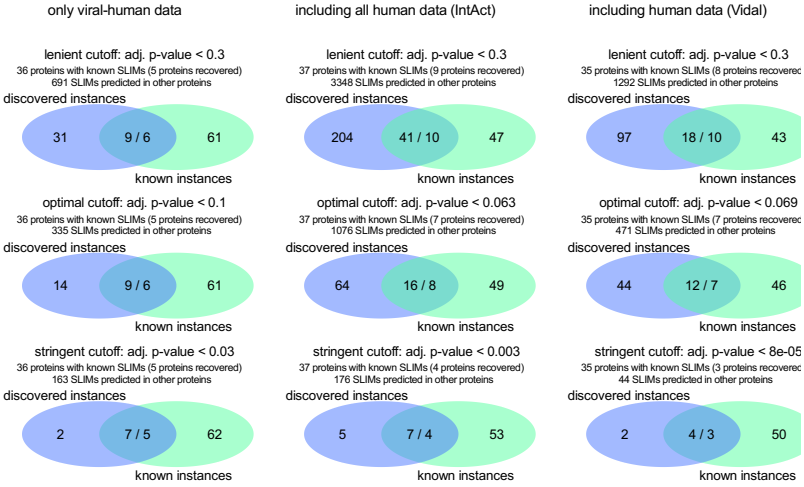
Supplementary figure 1



Supplementary figure 2



Supplementary figure 3



Supplementary figure 4