Surname/Family Name	:	Idrees
Given Name	:	Sobia
Abbreviation for degree as give in the University calendar	:	PhD
Faculty	:	Science
School	:	School of Biotechnology and Biomolecular Sciences
Thesis Title	:	Predicting motif mimicry in viruses

Abstract

One of the main pursuits in proteomics is to understand the complex network of protein-protein interactions (PPI) that underpin biological processes. Two major classes of PPI are domain-domain interactions (DDI) between globular proteins, and domain-motif interactions (DMI) between a globular domain and a short linear motif (SLiM) in its partner. Advances in high-throughput experimental techniques have been applied at large-scale in an attempt to characterise the interactomes of various organisms. However, the PPI networks identified by these high-throughput experiments have low resolution as compared to low-throughput technologies, such as protein co-crystallization. Furthermore, large-scale approaches may be poor at capturing low affinity or transient interactions, which includes the majority of known DMI. To date, several studies have been conducted to identify how well these PPI data can capture protein complexes, but the ability of high-throughput PPI-detection methods to capture DMI remains a largely unanswered question.

Here, a new computational pipeline (SLiMEnrich) was designed to assess how well a given source of PPI data captures DMIs and thus, by inference, how useful that data should be for SLiM discovery. To help system biologists choose appropriate methods for predicting different types of interactions, a comparison study of existing high-throughput PPI datasets was performed. PPI data, SLiM predictions, domain composition and known SLiM-domain binding partnerships were integrated to identify possible DMI and DDI within interactomes. SLiMEnrich identified PPI data that were enriched for DMI or DDI by randomising the PPI within the network to generate a background expectation. Moreover, it was found that host-pathogen PPI data can be used to study molecular mimicry in viruses and to discover novel SLiMs. An *in-silico* peptide exchange approach was developed and applied to provide additional validation of predicted mimicry candidates. Despite limitations of this technique in large-scale validation of predicted SLiMs and DMIs, peptide exchange simulations identified a few high-confidence SLiMs that are likely to bind known structures and therefore constitute strong candidates for molecular mimicry by human viruses.

COPYRIGHT STATEMENT

Signed ……………………………………………..........................

Date ……………………………………………...........................

AUTHENTICITY STATEMENT

Signed ……………………………………………..........................

Date ………………………………………………………………………

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The candidate contributed greater than 50% of the content in the publication and is the "primary author", ie. the candidate was responsible primarily for the planning, execution and preparation of the work for publication
- The candidate has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not:

□ This thesis contains no publications, either published or submitted for publication
(if this box is checked, you may delete all the material on page 2)

□ Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement
(if this box is checked, you may delete all the material on page 2)

□ This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the UNSW Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Candidate's Name	**Signature**	**Date (dd/mm/yy)**

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed ……………………………………………..............

Date ……………………………………………..............

Predicting motif mimicry in viruses

Sobia Idrees

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

UNSW AUSTRALIA

School of Biotechnology and Biomolecular Sciences

Faculty of Science

May 2019

Abstract

One of the main pursuits in proteomics is to understand the complex network of protein-protein interactions (PPI) that underpin biological processes. Two major classes of PPI are domain-domain interactions (DDI) between globular proteins, and domain-motif interactions (DMI) between a globular domain and a short linear motif (SLiM) in its partner. Advances in high-throughput experimental techniques have been applied at large-scale in an attempt to characterise the interactomes of various organisms. However, the PPI networks identified by these high-throughput experiments have low resolution as compared to low-throughput technologies, such as protein co-crystallization. Furthermore, large-scale approaches may be poor at capturing low affinity or transient interactions, which includes the majority of known DMI. To date, several studies have been conducted to identify how well these PPI data can capture protein complexes, but the ability of high-throughput PPI-detection methods to capture DMI remains a largely unanswered question. Here, a new computational pipeline (SLiMEnrich) was designed to assess how well a given source of PPI data captures DMIs and thus, by inference, how useful that data should be for SLiM discovery. To help system biologists choose appropriate methods for predicting different types of interactions, a comparison study of existing high-throughput PPI datasets was performed. PPI data, SLiM predictions, domain composition and known SLiM-domain binding partnerships were integrated to identify possible DMI and DDI within interactomes. SLiMEnrich identified PPI data that were enriched for DMI or DDI by randomising the PPI within the network to generate a background expectation. Moreover, it was found that host-pathogen PPI data can be used to study molecular mimicry in viruses and to discover novel SLiMs. An in-silico peptide exchange approach was developed and applied to provide additional validation of predicted mimicry candidates. Despite limitations of this technique in large-scale validation of predicted SLiMs and DMIs, peptide exchange simulations identified a few high confidence SLiMs that are likely to bind known structures and therefore constitute strong candidates for molecular mimicry by human viruses.

Table of Contents

Contribution made by others

Richard J. Edwards helped in designing and coding the SLiMEnrich application, wrote part of published Chapter 2 text, and helped in making Figure 2.1. Åsa Pérez-Bercoff helped in reviewing drafts of Chapter 2.

Acknowledgements

My sincerest thanks to my research supervisor Dr. Richard J Edwards, School of Biotechnology and Biomolecular Sciences, University of New South Wales for his immense knowledge, keen interest, guidance, encouraging attitude and patience throughput the last 3 years that enabled me to carry out my work.

No Acknowledgements could ever adequately express my feelings for my parents who always wished to see me walking ahead taking every step leading towards new achievements, and whose hand always raised with unlimited affection by virtue of which I could reach at this position.

Cordial and sincere obligations are rendered to my brothers and sisters for always supporting me.

I would also like to thank my friends in Sydney for making my time enjoyable and for always encouraging me throughout my academic period.

Finally, I would like to acknowledge the University of New South Wales for providing me University International Postgraduate Award to continue my PhD.

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
AP-MS	Affinity Purification coupled Mass Spectrometry
CLV Class	Cleavage Sites
CoFrac-MS	CoFraction coupled Mass Spectrometry
DEG Class	Degeneration Sites
DMIs	Domain Motif Interactions
DOC	Docking Sites
dProtein	Domain-containing Protein
d.p.	Decimal point
EBD	ELM Binding Domain
ELM	Eukaryotic Linear Motif
ELMdb	ELM Database
ELMi	ELM interactions
ELMc	ELM classes
ELMOcc	ELM Occurrences
E-score	Enrichment score
hPPIs	Human protein-protein interactions
hProt/hProtein	Human Protein
IDRs	Intrinsically Disordered Regions
LIG Class	Ligand Binding Sites
MOD Class	Post-translational Modification Sites
mProtein	Motif-containing Protein
PPIs	Protein-Protein Interactions
PTMs	Post-translational Modifications
s.f.	Significant figure
SLiMs	Short Linear Motifs
TRG Class	Targeting Sites
vhPPI	Virus host Protein-Protein Interactions
Y2H	Yeast two Hybrid

1 Chapter 1: Introduction

Short linear motifs (SLiMs) are linear recurring functional peptides/microdomains consisting of 2-15 contiguous residues (D'haeseleer 2006; Davey, Van Roey et al. 2012; Weatheritt, Davey et al. 2012; Bhowmick, Guharoy et al. 2015). SLiMs are important because their key residues are involved in the fundamental cellular processes including post-translational modifications (PTMs), sub-cellular localization, protein trafficking, regulatory functions, controlling cell cycle, signal transduction and stabilizing scaffolding (Dinkel and Sticht 2007; Van Roey, Uyar et al. 2014). However, SLiMs usually have only 2-5 defined positions which make them difficult to identify through computational as well as experimental methods (Neduva and Russell 2006; Gibson 2009; Pancsa and Fuxreiter 2012). The focus of this thesis was to explore methods to tackle these challenges and to apply them to study molecular mimicry in viruses through SLiMs. This chapter will first give an overview of SLiMs, their role in establishing protein-protein interactions, before introducing the concepts of molecular mimicry in human viruses.

1.1 Short linear motifs (SLiMs)

Proteins are primary function molecules of the cell. Many proteins carry out their functions through adapting a well-defined three-dimensional (3D) structure, but there are portions in the proteome which are not well defined yet are significant for carrying out different cellular functions. These unstructured regions are known as intrinsically disordered regions (IDRs) (van der Lee, Buljan et al. 2014). IDRs often contain SLiMs, which mediate interactions with other protein partners (1.2.2). These interactions are often transient and low affinity (1–150 μM range), which makes SLiMs particularly good at mediating functions that require fast response (Diella, Haslam et al. 2008). IDRs and SLiMs are enriched in alternative exons which make them significant in terms of functional diversity (Tompa 2012; Weatheritt, Davey et al. 2012).

The leading repository for curated data on SLiMs is the Eukaryotic Linear Motif (ELM) database (Dinkel, Michael et al. 2012; Dinkel, Van Roey et al. 2016). ELM identifies six main classes of motifs: ligand binding sites (LIG), cleavage sites (CLV), subcellular targeting sites (TRG), sites of PTMs (MOD), docking sites (DOC) and degradation sites (DEG).

SLiMs not only provide significant biological knowledge related to different cellular processes (Seo and Kim 2018), but also help in improving our understanding about the complexity of the interactome (Davey, Van Roey et al. 2012; Seo and Kim 2018). According to the recent studies, it has been seen that SLiMs residing in IDRs help in controlling specificity and range of the phospho-signalling which is regulated by anchoring proteins (Langeberg and Scott 2015; Nygren and Scott 2015). For example, protein kinase A (PKA) phosphorylation is regulated by the anchoring protein, namely A-kinase anchoring protein (AKAP) (Welch, Jones et al. 2010). AKAPs have the conformational flexibility which makes them ideal for adapting to the signalling requirements. This is the reason that these proteins have ability to perform different functions at different locations of the cell. According to one hypothesis, AKAPs are likely to create signalling domains that are spatially constrained, which makes PKA/AKAP as an ideal drug target (Nygren and Scott 2015).

SLiMs are not only important for critical cellular processes but can also lead to serious health issues if anything goes wrong with them (Ward, Sodhi et al. 2004). SLiMs are now being considered as an ideal drug targets that can help cope with the dreadful diseases (Uyar, Weatheritt et al. 2014). According to a recent analysis, 22% of the mutations in human proteome occur in the unstructured region which suggests SLiMs as important players in diseases (Ward, Sodhi et al. 2004). This is why SLiMs are being investigated for their role in diseases, to define potential therapeutics especially against viral infections (Neduva and Russell 2006).

1.1.1 SLiM evolution

The evolution of SLiMs can be explained by two main principles, namely divergent evolution and convergent evolution. Conservation of individual SLiM instances in proteins (homology) is known as divergent evolution while the independent evolution of SLiM instances in unrelated proteins is known as convergent evolution (Davey, Van Roey et al. 2012). SLiMs generally have higher evolutionary plasticity than other structurally or functionally constrained residues. A single point mutation can alter the SLiM occurrence through either destroying its functionality or creating a functionally active SLiM from an inactive protein sequence. This higher level of plasticity helps in rapid rewiring of PPI networks through establishing different SLiM mediated interactions (Neduva and Russell 2005).

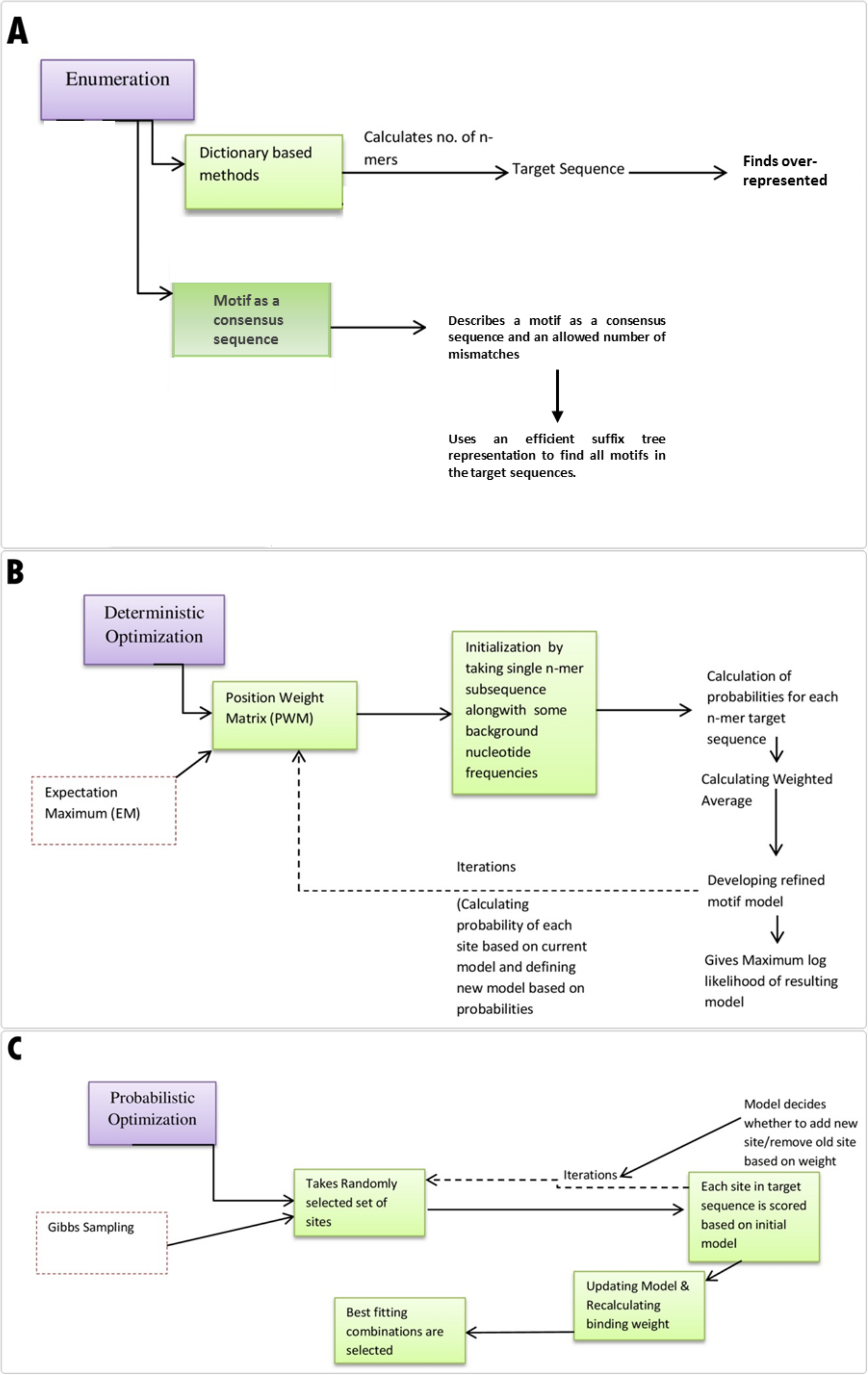
SLiMs are generally conserved, but are not as conserved as domains, which can make it difficult to find them without targeted methods (Nguyen Ba, Yeh et al. 2012; Bhowmick, Guharoy et al. 2015; Edwards and Palopoli 2015). However, their convergent evolution means that they are not necessarily conserved and imparts evolutionary plasticity with all its important implications (Edwards and Palopoli 2015). SLiMs mostly occur in the IDRs which give them more versatility and the ability to interact with different partners. These features of SLiMs show that these linear motifs are prone to have independent origins, and these can help finding novel motifs that share interaction partners (Neduva and Russell 2005). In general, this SLiM plasticity creates an Achilles heel which helps pathogenic proteins imitate host proteins and help them to interact with host cellular pathway (Davey, Trave et al. 2011).

1.1.2 SLiM discovery algorithms and tools

SLiM prediction methods are continually evolving to ease the process of SLiM discovery. To date, several algorithms have been developed to predict motifs with minimal false positives. These methods either tend to discover new motifs (*de-novo* discovery) or help finding motif instances of already available data. There are three main categories of the motif discovery algorithms: deterministic optimisation, enumeration and probabilistic optimization **(Table 1.1)** (D'haeseleer 2006).

Table 1.1. Main types of SLiM discovery algorithms.

Algorithm	Description	Advantages	Disadvantages	Implementation examples
Enumeration	Covers the space for all possible motifs for a specific model such as dictionary-based methods	No Risk to get stuck in local optimum	-Sometimes too rigid -May overlook some of the subtle patterns available in actual binding sites	1. Dyad-Analysis (van Helden, Rios et al. 2000) 2. YMF (Sinha and Tompa 2000) 3. MOPAC (Ganesh, Siegele et al. 2003) 4. DMotif (Sinha 2003) 5. MaMF (Hon and Jain 2006)
Deterministic Optimization	This works based on the Expectation Maximization (EM) and generates a position weight matrix (PWM)	No Risk to get stuck in poor local maximum	Covers small subset of the known binding sites	1. LOGOS (Xing, Wu et al. 2004) 2. PhyME (Sinha, Blanchette et al. 2004) 3. OrthoMEME (Prakash, Blanchette et al. 2004) 4. ALSE (Leung and Chin 2006)
Probabilistic Optimization	This works based on Gibbs Sampling and uses a weighted sample from sub sequences	-Highly focused on best fitting combinations -Can detect subtle block-based motifs	Lack of accuracy for unrelated proteins	1. MotifSampler (Thijs, Marchal et al. 2002) 2. PhyloGibbs (Siddharthan, Siggia et al. 2005)


A
Enumeration
Dictionary based methods
Calculates no. of n-mers
Target Sequence
Finds over-represented
Motif as a consensus sequence
Describes a motif as a consensus sequence and an allowed number of mismatches
Uses an efficient suffix tree representation to find all motifs in the target sequences.
B
Deterministic Optimization
Position Weight Matrix (PWM)
Expectation Maximum (EM)
Initialization by taking single n-mer subsequence alongwith some background nucleotide frequencies
Calculation of probabilities for each n-mer target sequence
Calculating Weighted Average
Developing refined motif model
Gives Maximum log likelihood of resulting model
Iterations
(Calculating probability of each site based on current model and defining new model based on probabilities
C
Probabilistic Optimization
Gibbs Sampling
Takes Randomly selected set of sites
Iterations
Model decides whether to add new site/remove old site based on weight
Each site in target sequence is scored based on initial model
Updating Model & Recalculating binding weight
Best fitting combinations are selected

(legend on next page)

Figure 1.1. Schematic diagram of SLiM discovery algorithms.

A) The Enumeration algorithm works in different ways for motif discovery. First one is the dictionary-based method where no. of n-mers is calculated in the target sequence and then finds the over-represented motifs. Alternatively, it describes the motif as a consensus sequence with allowed no. of mismatches and tree representation is used to find all the motifs. **B)** Deterministic Optimization is based on the Expectation Maximization (EM) that creates a position weight matrix (PWM). The process starts by taking single n-mer subsequence with some background residue frequencies. Then the probabilities of each n-mer sequences are calculated followed by weighted average. Based on this weighted average, a refined model is developed. This process is iterated to find the probability of each site and then the maximum log likelihood of resulting model is generated. **C)** Probabilistic Optimization is based on Gibbs Sampling which takes randomly selected set of sites and score them according to the initial model. This process keeps iterating and model decides whether to add new sites or remove old site based on the weight. After that, the model is updated, and binding weights are recalculated. Then the best fitting combinations are selected based on binding weights **(D'haeseleer 2006)**.

Development of new Bioinformatics tools for SLiM prediction has always been challenging because of the possibility of high false discovery rate. One of the challenges faced during the development of SLiM prediction tools is the robustness of the benchmarking. An adequate amount of benchmarking data is required to test new methods to see if they are working better than the existing methods (Edwards and Palopoli 2015). During the past few years, significant progress in the field of motif discovery has been seen and different new Bioinformatics tools have been developed to predict SLiMs from sequence data. There are three main categories of SLiM prediction tools, namely SLiM discovery from known motifs, *de-novo* SLiM discovery and user-defined motif tools. Most of the SLiM discovery tools use known motifs to look for new instances in protein sequences e.g. SLiMProb (Davey, Haslam et al. 2011), ScanSite (Obenauer, Cantley et al. 2003) and iELM (Weatheritt, Jehl et al. 2012), while tools such as QSLiMFinder (Palopoli, Lythgow et al. 2015) and motif-x (Chou and Schwartz 2011) look for completely new motifs. There are also tools which take predefined motifs/information to predict new motifs such as 3of5 (Seiler, Mehrle et al. 2006) and SLiMSearch (Davey, Haslam et al. 2011). All of these tools are facilitating discovery of new SLiMs to help expand current knowledge of linear motifs in the proteome **(Table 1.2)**.

Table 1.2. List of computational tools used for SLiM prediction.

	Tool	Description	Availability	Reference
SLiM discovery tools based on known motifs	PROSITE	PROSITE was the first catalogue of the linear motifs. This database is now focusing on protein signatures and globular domains	http://prosite.expasy.org/	(Bairoch 1993)
	Scansite	Scansite predicts motifs that are important for cell signalling. It uses profile-based searches of known motifs against user sequences.	http://scansite.mit.edu.	(Obenauer, Cantley et al. 2003)
	ScanProsite	Scanprosite uses regex searches to find motifs against protein sequences either defined by users or in public databases.	http://prosite.expasy.org/scanprosite/	(de Castro, Sigrist et al. 2006)
	SLiMProb (Short Linear Motif Probability)	Searches user defined motifs against local protein data.	http://www.slimsuite.unsw.edu.au/servers.php	(Davey, Haslam et al. 2011)
	MnM (Minimotif Miner)	Identifies SLiMs in a protein sequence that have known function in some other protein.	http://mnm.engr.uconn.edu	(Balla, Thapar et al. 2006)
	iSPOT (Infer Sequence Prediction of Target)	Uses structural data to predict SH3, PDZ and WW binding sequences	http://cbm.bio.uniroma2.it/ispot	(Brannetti and Helmer-Citterich 2003)
	ELM Database (Eukaryotic Linear Motif Database)	This database contains manually curated and experimentally validated SLiMs of eukaryotes. It is one of the biggest resource to analyse functional SLiMs.	http://elm.eu.org	(Dinkel, Van Roey et al. 2014)
	iELM (Interactions of Eukaryotic Linear Motif)	Uses PPI data to predict instances of known motifs	http://i.elm.eu.org	(Weatheritt, Jehl et al. 2012)
	AMS (AutoMotifServer)	This predicts motifs using a trained support vector machine (SVM). This is used for prediction of PTMs.	http://code.google.com/p/automotifserver/	(Plewczynski, Basu et al. 2012)

	Tool	Description	Availability	Reference
***De novo* SLiM Discovery Tools**	qPMS7	Uses LDMS patterns without correcting homology	http://pms.engr.uconn.edu/downloads/qPMS7.zip	(Dinh, Rajasekaran et al. 2012)
	Pratt	Predicts motifs based on over representation regex without homology correction	http://www.ebi.ac.uk/Tools/pfa/pratt/	(Jonassen, Collins et al. 1995)
	SLIDER (LDMS CMM tool)	Maps motifs on PPI interfaces to find correlated motifs.	http://bioinformatics.uhasselt.be	(Boyen, Van Dyck et al. 2011)
	TEIRESIAS	Uses text patterns to search for motifs	http://code.google.com/p/teiresias	(Rigoutsos and Floratos 1998)
	SLiMMaker (Short Linear Motif Maker)	Align peptide sequences and generates regex consensus sequences	http://www.slimsuite.unsw.edu.au/servers.php	(Palopoli, Lythgow et al. 2015)
	NestedMICA (Nested Motif Independent Component Analysis)	Identifies enriched motifs against reference proteins	http://www.sanger.ac.uk/Software/analysis/nmica/	(Dogruel, Down et al. 2008)
	ANCHOR	Uses user defined regex and maps them on disorder profiles	http://anchor.enzim.hu	(Dosztanyi, Meszaros et al. 2009)
	PepSite	Uses structural data for prediction of DMI	http://pepsite2.russelllab.org/	(Dosztanyi, Meszaros et al. 2009)
	MoRFpred (MoRF predictor)	Finds regions within IDR based on the propensity of order	http://biomine.ece.ualberta.ca/MoRFpred/	(Disfani, Hsu et al. 2012)
	MotifCluster	Uses PPI data to find correlated motifs	http://bmf.colorado.edu/motifcluster	(Leung, Siu et al. 2009)
	D-MIST (Domain-Motif Interaction from Structural Topology)	Uses structural context to predict DMI from PDB	N/A	(Betel, Breitkreuz et al. 2007)

	Tool	Description	Availability	Reference
	SLiMScape (SLiM plugin for Cytoscape)	A plugin designed for Cytoscape to analyse SLiMs	http://apps.cytoscape.org/apps/slimscape	(O'Brien, Haslam et al. 2013)
	D-MOTIF (LDMS CMM tool)	Uses PPI data to find correlated motifs	http://meme-suite.org/	(Bailey and Gribskov 1997)
	SLiMFinder Short Linear Motif Finder)	This tool helps finding over represented motifs in unrelated proteins.	http://www.slimsuite.unsw.edu.au/servers.php	(Edwards, Davey et al. 2007)
	QSLiMFinder (Query SLiMFinder)	Query based SLiM discovery tool with better sensitivity and specificity.	http://www.slimsuite.unsw.edu.au/servers.php	(Palopoli, Lythgow et al. 2015)
	SLiMDisc (Short Linear Motif Discovery)	This tool has a heuristic approach where over-represented motifs are ranked in unrelated proteins. This was one of the first de novo tool capable of SLiM predictions.	http://bioware.ucd.ie/	(Davey, Shields et al. 2006)
	MFSPSSMPred (Masked, Filtered and Smoothed Position Specific Scoring Matrix based Predictor)	Predicts motifs based on evolutionary conservation or sequence features.	http://biomine-ws.ece.ualberta.ca/MoRFpred/index.html	(Fang, Noguchi et al. 2013)
	SLiMPrints (Short Linear Motif Fingerprints)	Uses statistical models to find conservation fingerprints	http://bioware.ucd.ie/	(Davey, Cowan et al. 2012)
	SLiMPred (Short Linear Motif Predictor)	Prediction of SLiMs in protein sequences	http://bioware.ucd.ie/	(Mooney, Pollastri et al. 2012)
	MEME (Multiple Em for Motif Elicitation)	Uses EM to find DNA/protein motifs	http://meme.nbcr.net	(Bailey and Elkan 1994)
	FIRE-pro (finding Informative Regulatory Elements in proteins)	Identifies correlated motifs using mutual information	https://tavazoielab.c2b2.columbia.edu/FIRE-pro/	(Lieber, Elemento et al. 2010)

	Tool	Description	Availability	Reference
***De novo* SLiM Discovery Tools**	D-STAR (LDMS CMM tool)	Uses PPI data to identify correlated motifs	N/A	(Tan, Hugo et al. 2006)
	DILIMOT (DIscovery of Linear MOTifs	Prediction of motifs that are over represented in a set of proteins that interact with the target protein	http://dilimot.russelllab.org/	(Neduva and Russell 2006)
	motif-x	Uses over represented peptides in combination with background amino acid frequencies to generate fixed position motifs	http://motif-x.med.harvard.edu/	(Chou and Schwartz 2011)
	MOTIPS (MOTIf analysis pipeline)	Uses short peptides in combination with DMI to predict over represented profiles	http://motips.gersteinlab.org/	(Lam, Kim et al. 2010)
	GLAM2 (Gapped Local Alignment of Motifs)	Uses Gibbs Sampling and simulated annealing to find over represented patterns	http://bioinformatics.org.au/glam2	(Frith, Saunders et al. 2008)
User defined tools	FIMO (Find Individual Motif Occurrences)	MEME profiles are searched against public databases/user defined proteins	http://meme.sdsc.edu	(Grant, Bailey et al. 2011)
	PRESTO (Protein Regular Expression Search Tool)	Uses regex searches to find SLiMs against local protein data.	http://slimsuite.blogspot.com.au/	(Edwards 2013)
	3of5 (3of5 regex search tool)	Uses regex searches to find sequences	http://dkfz.de/mga2/3of5/3of5.html	(Seiler, Mehrle et al. 2006)
	MAST (Motif Alignment and Search Tool)	Uses multiple profile motifs for identification	http://meme-suite.org/	(Bailey and Gribskov 1997)
	SLiMSearch 2.0	Proteome wide searches to find predefined motifs	http://bioware.ucd.ie/	(Davey, Haslam et al. 2011)

1.2 Protein-protein interactions (PPIs)

Protein-protein interactions (PPIs) are known to mediate diverse functions such as catalysis of metabolic reactions, transportation of molecules, modification of kinetic properties of enzymes and altering specificity of proteins (De Las Rivas and Fontanillo 2012). During signalling events, different proteins interact with each other to maintain cell growth and other cellular processes through pathway regulation. During the past years, different studies have been conducted to discover/predict complete maps of PPIs in different organisms (Rajagopala, Sikorski et al. 2014; Rolland, Tasan et al. 2014; Hein, Hubner et al. 2015; Zhang, Ou-Yang et al. 2015). This knowledge is being widely used to get insights into cellular organization of the organism as well as to cure different diseases such as cancer, viral and bacterial infections through targeting PPIs and disrupting signalling events (Seo and Kim 2018).

PPIs are either regarded as direct or indirect interactions. Direct interactions are established when two proteins physically interact with each other. On the other hand, indirect interactions are established when two proteins interact in the presence of an intermediate protein which leads to the formation of complexes (Peng, Wang et al. 2017). Another important feature of PPIs is their nature of interactions which can either be transient or permanent interactions based on stability and lifetime (Bhowmick, Guharoy et al. 2015). Most of the time permanent interactions are long term and result into stable complexes. For example, complexes such as RNA polymerase and haemoglobin are assembled by proteins with stable interactions (Peng, Wang et al. 2017). On the other hand, transient interactions have shorter lifetime as they associate and dissociate quickly and tend to happen under certain biological contexts. Such interactions enables cells to respond quickly to extracellular stimuli (Lubovac, Gamalielsson et al. 2006; Seo and Kim 2018). Stable and permanent interactions are mostly mediated through domains (DDIs) and transient interactions are mostly mediated by a motif in one protein and domain in other protein (DMIs) (Bhowmick, Guharoy et al. 2015) **(Figure 1.2),** which are discussed in more detail in the following sections.

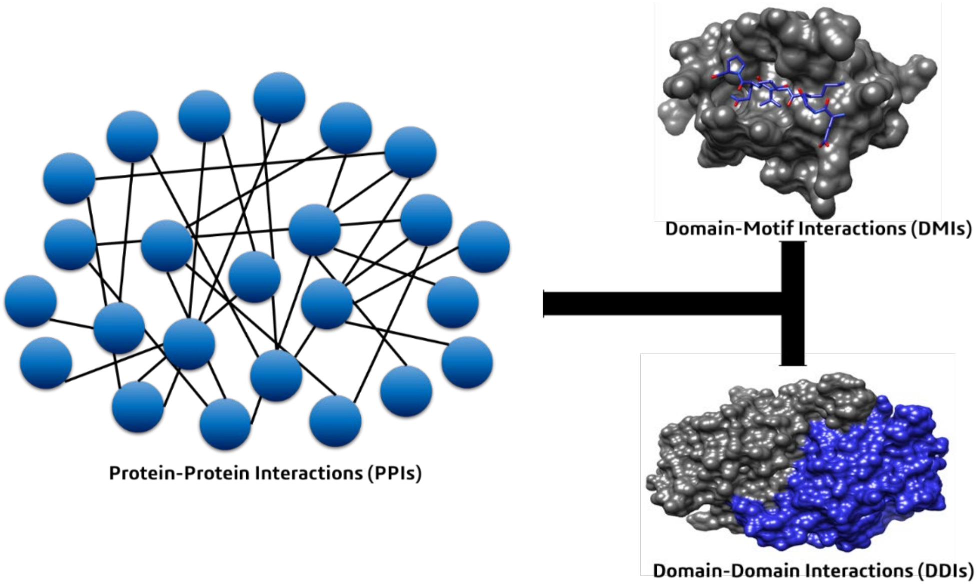

Domain-Motif Interactions (DMIs)
Protein-Protein Interactions (PPIs)
Domain-Domain Interactions (DDIs)

Figure 1.2. Types of Protein-protein Interactions.

An illustration of Domain-motif interactions (DMIs) where motif in one protein (shown in blue) interacts with domain of other protein (shown in grey), Domain-domain interactions (DDIs) where a domain in one protein (shown in in blue) interacts with a domain in other protein (shown in grey).

1.2.1 Domain-domain interactions (DDIs)

Most known PPIs are domain-domain interactions (DDIs) mediated by globular domains in different proteins. These DDIs involve large interfaces between protein domains (Diella, Haslam et al. 2008). Most of the interaction data generated by PPI detection experiments has a possibility of false positives and false negatives. Studying DDIs where a domain in one protein interacts with a domain in other protein can deal with these limitations and can provide important clues to understand the intricacy of biological systems (Kim, Min et al. 2012). DDIs can be identified based on the 3-dimensional structures of protein complexes available in Protein Data Bank (Rose, Prlic et al. 2017). Different databases such as iPfam (Finn, Miller et al. 2014) and 3DID (Mosca, Ceol et al. 2014) have extracted DDIs from the known 3D structures. The one problem concerned with such databases is their number of interactions due to insufficient known 3D structures of proteins. This is the reason several computational methods have been developed to predict DDIs though there is not a single platform to integrate predicted DDIs by these methods. During recent years, two databases such as DOMINE (Raghavachari, Tasneem et al. 2008) and UniDomInt (Bjorkholm and Sonnhammer 2009) have been specifically developed to store DDIs from different resources. One advantage of these databases is their confidence score which gives reliability of the predicted DDIs. Despite of providing significant DDIs, these databases are outdated and do not contain any recently published datasets.

1.2.2 Domain-motif interactions (DMIs)

Another mode of interaction is through domain-motif interactions (DMIs) which are mediated by SLiMs (D'haeseleer 2006; Davey, Van Roey et al. 2012; Weatheritt, Davey et al. 2012; Bhowmick, Guharoy et al. 2015). These DMIs are basically a subset of PPIs where a protein structure is induced in a SLiM of other protein (Dinkel and Sticht 2007; Gibson 2009; Pancsa and Fuxreiter 2012; Van Roey, Uyar et al. 2014). DMIs are often transient in nature and are known to be involved in different signalling processes including protein targeting and signal transduction.(Pawson, Raina et al. 2002). Specific SLiMs interact with specific domains to establish a DMI (e.g. proline rich motifs tend to interact with SH3 domains) (Kaneko, Li et al. 2008).

SLiMs usually have 2-5 conserved positions that are essential to interact with their partner domains while other positions are less conserved. This flexibility in the sequence pattern helps in establishing different DMIs i.e. a single motif can bind to several domains from same family or variants of same motif can bind with the same domain. For example, PDZ domains are known to interact with variants of same motif (class I (x[S/T]xΨ-COOH), class II (xΨxΨ-COOH) and class III (x[E/D]xΨ-COOH)) (Nourry, Grant et al. 2003). This characteristic of DMIs makes them promiscuous in nature though they also show specificity of binding. The specificity is often dependent on the sequence context of motif which serves as a scaffold for establishing DMIs while contextual residues help in defining interaction specificity (Seet, Dikic et al. 2006; Miller, Jensen et al. 2008; Stein and Aloy 2008; Akiva, Friedlander et al. 2012). The transient nature of DMIs make it challenging to capture them through high-throughput screens and therefore new computational as well as experimental methods are much needed to predict and validate DMIs (Seo and Kim 2018).

1.3 High-throughput methods to detect PPIs

One of the main pursuits in proteomics is to understand the organization of PPIs as a complex network. To date, several studies have been conducted to identify PPIs, but most of these are detected by small scale experiments. During recent years, different high-throughput methods have been developed to detect large number of PPIs with reliability such as Affinity Purification coupled Mass Spectrometry (AP-MS) and Yeast two hybrid (Y2H) (Blikstad and Ivarsson 2015). Recent advances in high-throughput experimental techniques have led to large amount of PPI data which is providing rough picture of how two proteins interact in biological system. However, PPI networks being identified by these high-throughput experiments have low resolution as compared to PPIs from low-throughput technologies such as protein co-crystallization. Another problem being faced in terms of high-throughput PPI data is their relatively high error rates and protocol specific biasness (Seo and Kim 2018). Moreover, there aren't enough experimental evidences to show how good a method is. New computational as well as experimental methods are much needed to study mechanism of interactions among two proteins. The PPIs identified by these methods are modeled as a network where proteins are called nodes and interactions among them are regarded as edges. The data generated by PPI detection methods is being used to study biological pathways, protein complexes, protein functionality and to identify potential drug targets. However, to ensure that the knowledge gained by studying PPIs is biological meaningful, it is important to ensure the quality of the detected PPIs (Stein and Aloy 2008; Kim, Sabharwal et al. 2010).

Each technique has their own advantages and disadvantages. Small scale experiments can detect low numbers of PPIs, though their quality of interactions is often high. On the other hand, high-throughput methods can detect a large number of PPIs, but the interaction quality is often low.

Despite the efficiency of these high-throughput experiments, there is always the possibility of false negatives and false positives which increases troubles for successful predictions (Li, Wu et al. 2010; Zhang, Lin et al. 2015). Studies have shown that binary PPI detection methods are more susceptible to get false positive interactions (Rajagopala, Sikorski et al. 2014). Different factors including poor expression, cofactors, binding partners and lack of necessary posttranslational modifications are considered responsible for high false discovery rate (Peng, Wang et al. 2017). Another limitation of this approach is that proteins are often overexpressed which can lead to non-specific interactions, raising overall false positive rate (Blikstad and Ivarsson 2015). On the other hand, co-complex methods, including AP-MS and CoFrac-MS are susceptible to contamination by abundant proteins which are co-purified from the pull down (Zhang, Lin et al. 2015). Another major limitation of this approach is that it cannot detect weak or transient interactions (Peng, Wang et al. 2017). This is the reason that experimental as well as computational methods to validate these interactions are needed. One problem often faced to detect PPIs is their physiological settings during the experiment as certain PPIs occur at certain conditions. Moreover, several factors can also influence PPI detection such as transient nature, PTMs, abundance of proteins and IDRs. It can be said that unravelling the proteome wide interactome is quite challenging. The two most common ways to evaluate reliability of PPIs is to design new methods to validate the interactions or to develop new computational methods to assess reliability of interactions through filtering out possible false positives from the data by finding the probability of the observed PPI (Pitre, Alamgir et al. 2008; Kim, Sabharwal et al. 2010).

1.3.1 Yeast two hybrid (Y2H)

Yeast two hybrid (Y2H) is a well-known technique to detect PPIs in yeast cells. The interacting proteins are called bait and prey which interaction activates reporter genes that results into a color reaction or growth on specific media **(Figure 1.3).** Y2H is being used to identify genome-wide interactions in different organisms (Bruckner, Polge et al. 2009) namely humans (Rolland, Tasan et al. 2014), *Caenorhabditis elegans* (Simonis, Rual et al. 2009), bacteriophage T7 (Hauser, Blasche et al. 2012), *Drosophila melanogaster* (Formstecher, Aresta et al. 2005) *and Saccharomyces cerevisiae* (Yu, Braun et al. 2008). Y2H is considered a powerful systems biology tool to study large interactomes or to understand diseases through understanding mechanisms of protein interactions in a system (Lim, Hao et al. 2006).

There are two screening approaches: the library approach and the array approach. In library approach, pairwise interactions are searched between protein of interests (bait and prey) available in cDNA libraries. The one disadvantage of this approach is the possibility of false positives (rate of wrongly identified proteins). It requires colony PCR and sequencing techniques to identify interaction partners, which makes this approach time consuming and costly. On the other hand, array also known as matrix approach identifies interactions through direct mating a pool of baits with a pool of preys in different yeast mating types. The advantage of this approach is its automated nature that can be used to identify genome wide interactomes. The disadvantage of this approach is it misses to identify certain interactions known as false negatives because of its restriction to have a limited set of full length open reading frames (ORFs) (Bruckner, Polge et al. 2009).

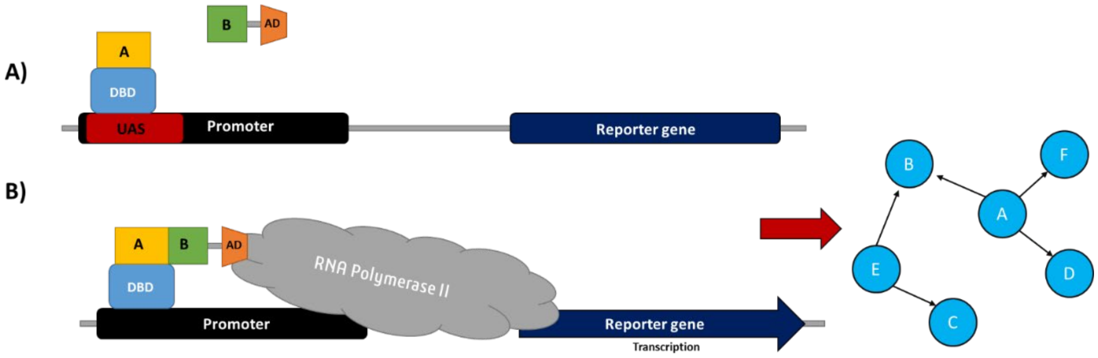


Figure 1.3. Yeast two hybrid approach.

A) Yeast two hybrid system requires a DNA binding domain (DBD) and an activation domain of a yeast transcription factor. A bait (protein of interest) is fused with the DBD and a prey (potential interacting protein) is fused with the AD.

B) The bait protein (A) upon fusion with the DBD, interacts with a binding site in the promoter region of a reporter gene. The prey protein (B) upon fusion with the AD, activates the gene expression through binding with the bait protein (A).

1.3.2 Affinity purification coupled mass spectrometry (AP-MS)

Affinity purification coupled mass spectrometry (AP-MS) has also become a powerful tool in systems biology to identify large-scale interactions. The technological advances in Mass spectrometry have revolutionized biochemical methods such as chemical cross-linking or affinity purification to detect proteome-wide interactions in different systems (Blikstad and Ivarsson 2015). In AP-MS technique, a protein is fused to a tag which is either detected by a specific antibody or an affinity column recognizing the tag **(Figure 1.4).** AP-MS approach can be used a single step purification where an individual tag known as Flag-tag is used to immunoprecipitated the protein. However, a two-step purification is often considered more efficient where proteins are either double tagged e.g. 6xHis- and Strep-tag or have two tags on either C- or N-terminal end of the protein which are separated by a cleavage site known as tandem affinity purification (TAP). This gives multiprotein complexes that contain the tagged protein. MS is then used to identify components of the complexes. Two step purification approach provides better sensitivity and specificity. Recent advances in MS analysis and computational methods have improved accuracy of PPI identification and validations (Bruckner, Polge et al. 2009).

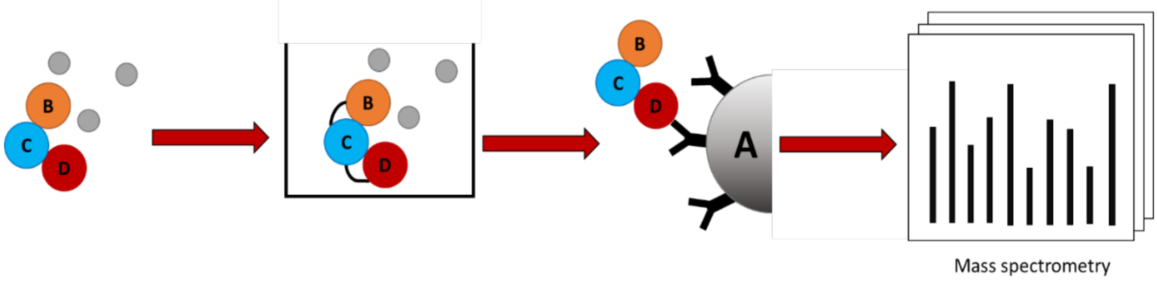


Figure 1.4. Affinity purification coupled mass spectrometry (AP-MS).

In AP-MS, protein of interest (A) is tagged with a tag protein. Proteins that bind to the tagged protein are co-purified. These proteins are then identified by mass spectrometry (MS).

1.3.3 Co-Fractionation followed by mass spectrometry (CoFrac-MS)

Co-fractionation followed by Mass Spectrometry (CoFrac-MS) has been found to be more suitable for identifying co complex interactions, including direct and indirect interactions between proteins (Kim, Sabharwal et al. 2010). In CoFrac-MS approach, protein extract is extensively fractioned using biochemical methods (e.g. size exclusion chromatography) which are then detected by MS **(Figure 1.5).** Just like AP-MS, this method can be used at proteome level, but it's often difficult to distinguish between direct or indirect interactions between protein pairs (Luck, Sheynkman et al. 2017).

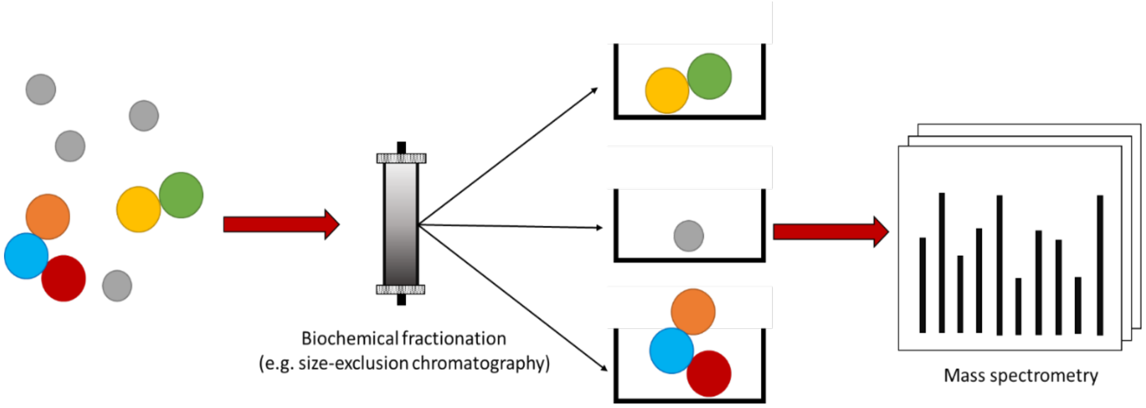

Biochemical fractionation
(e.g. size-exclusion chromatography)
Mass spectrometry

Figure 1.5. Co-Fractionation followed by Mass spectrometry.

In CoFrac-MS, extensive fractionation is done on protein extract to separate protein complexes which are then detected by MS.

1.4 Public protein-protein interaction repositories

Protein-protein interactions are vital for the proper functioning of the cells. According to an estimation, the number of PPIs is around 130,000-650,000 but still, the exact number of PPIs is unknown (Baspinar, Cukuroglu et al. 2014). These PPIs are being collected in specialized databases, allowing better analysis of the protein network. The first database created for maintaining PPI data was the Database of Interacting Proteins (DIP) (Xenarios, Rice et al. 2000). Public PPI repositories are growing rapidly. These repositories are not only managing PPI data, but are also helping in identifying novel SLiMs. To date, several PPI repositories have been developed and each one of them has their own advantages to study protein networks. There are three main categories of the PPI repositories i.e. primary databases which contain curated data directly from experiments or literature, meta-databases which contain curated data from experiments as well as from other PPI resources and prediction-based databases which contain data from experiments and different prediction methods. I selected a representative sample of PPI databases that seek to be comprehensive in coverage **(Table 1.3)**.

Table 1.3. Public protein-protein interaction repositories.

	Repository	Types of data	Species	Latest Release	Release Frequency	Availability
Primary Databases (Curation based on experimental data/Literature)	DIP (De Las Rivas and Fontanillo 2010) (Database of Interacting Proteins)	PPIs	All	Jan, 2014	No specific frequency	http://dip.doe-mbi.ucla.edu/
	BioGRID (Oughtred, Chatr-Aryamontri et al. 2016) (Biological General Repository for Interaction Datasets)	PPIs, Transcription data	All	v3.5.167 Nov, 2018	Monthly	http://wiki.thebiogrid.org/doku.php/statistics
	HPRD (Keshava Prasad, Goel et al. 2009) (Human Protein Reference Database)	PPIs, PTMs and subcellular localization	Human	Release 9 Apr, 2010	No recent updates	http://www.hprd.org/
	BIND (Keshava Prasad, Goel et al. 2009) (Biomolecular Interaction Network Database)	PPIs	All	Not active	N/A	http://bond.unleashedinformatics.com/
	MINT (Licata, Briganti et al. 2012) (Molecular INTeraction database)	PPIs, DNA/RNA interactions	All	Sep 2013	N/A	http://mint.bio.uniroma2.it/mint/
	IntAct (Orchard, Ammari et al. 2014)	PPIs	All	v2.0 Nov, 2018	Monthly	http://www.ebi.ac.uk/intact/
	DOMINO (Ceol, Chatr-aryamontri et al. 2007)	DMIs	Human, Mouse, Rat, Yeast	Oct, 2009	Not active	http://mint.bio.uniroma2.it/domino/
Meta databases (Curations based on experimental data and integration with other interaction	PINA (Cowley, Pinese et al. 2012) (**P**rotein **I**nteraction **N**etwork **A**nalysis)	PPIs	All	May, 2014	No specific frequency	http://cbg.garvan.unsw.edu.au/pina/
	APID (Prieto and De Las Rivas 2006)	PPIs	All	March 2018	3 months	http://bioinfow.dep.usal.es/apid/
	HINT (Das and Yu 2012) (*H*igh-quality *INT*eractomes)	PPIs	All	v4.0, Nov,2018	Nightly	http://hint.yulab.org/

	iRefWeb (Turner, Razick et al. 2010)	PPIs	All	v13, June, 2014	Yearly	http://wodaklab.org/iRefWeb/
	Cpdb (Kamburov, Stelzl et al. 2013) (ConsensusDB)	Different sorts of interactions	Human, Yeast, Mouse	v32, Jan, 2017	Yearly	http://consensuspathdb.org/
Prediction databases (Curations based on experimental and predicted data)	PIPs (McDowall, Scott et al. 2009) (Protein-protein interaction prediction)	PPIs	Human	Sep, 2008	No recent updates	http://www.compbio.dundee.ac.uk/www-pips/index.jsp
	OPHID (Brown and Jurisica 2005) (Online Predicted Human Interaction Database)	PPIs	Human	v2.9, Sep, 2015	Yearly	http://ophid.utoronto.ca
	UniHI (Kalathur, Pinto et al. 2014) (Unified Human Interactome)	PPIs	Human	Mar 2017	No specific frequency	http://www.unihi.org/
	STRING (Szklarczyk, Franceschini et al. 2015)	PPIs	All	v10.5, May 2017	Every 2 Years	http://string-db.org/

1.5 High throughput PPI experiments and SLiM-mediated interactions

DMIs are being studied to identify novel SLiMs. Unfortunately, most of the available DMI knowledge has been derived from low-throughput studies. During recent years, different high-throughput methods such as arrays of protein/peptide, affinity purification, yeast two hybrid and display of peptides on yeast/phage have been used to study DMIs in different organisms (Blikstad and Ivarsson 2015). These high-throughput methods have generated large set of PPI data which is being used to predict protein complexes as well as functional SLiMs. Despite the efficiency of these high-throughput experiments, there is always the possibility of false negatives and false positives which increases troubles for successful predictions (Li, Wu et al. 2010; Zhang, Lin et al. 2015).These high-throughput experiments have been applied to different domain families, giving a significant amount of PPI data. Nowadays, studies are being conducted to find SLiMs in conjunction with their binding partners in human proteome (Rajagopala, Sikorski et al. 2014). Availability of high-throughput PPI data is facilitating development of novel computational tools for SLiM predictions. These computational tools are not accurate and might result in false positives. This problem is now being resolved through including gene ontology (GO), gene expression as well as high-throughput data for the execution of these methods (Zhang, Lin et al. 2015). Available high-throughput experiments have three main categories such as display methods, arrays and protein-fragment complementation assays **(Table 1.4).**

Table 1.4. Comparative table of available high-throughput methods for capturing motif mediated interactions.

Category	Method	Description	Advantages	Disadvantages
Microarray	Peptide Array	Peptide arrays chemically synthesize peptides having known sequences.	Uses known peptide sequences	Possibility of false negatives and false positives
			Helps determining non-binding peptides	Biasness
			Incorporation of non-natural/modified amino acids.	Limited coverage
			Direct mapping of interactions regulated by posttranslational modifications.	High Cost
	Protein Array	The main principle of this method is the immobilization of the protein of interest on the surface and probing to a labelled peptide/protein.	Investigation of PPI at large scale and PPIs related to PTMs	Labour intensive set-up
			Sample consumption at lower level than other methods.	Stability of proteins
			Gives quantitative information	
Display Methods	Peptide Phage Display	Analyses peptide binding domains and their binding specificities.	Highly diverse peptide libraries	Availability of data analysis, and expression constructs.
			Low Cost	Not suitable for interactions related to PTMs
	Yeast Surface Display	This method uses Yeast cells that carry plasmid DNA encoding peptides. These peptides are displayed on the surface of the cells.	Availability of information on non-binding clones	Lower throughput than phage display
			Can help investigate PPIs to some extent	
	Y2H	This method splits a transcription factor binding domain and a DNA binding domain to a bait protein.	Helps characterizing peptide binding motifs.	The higher possibility of false positives and false negatives
			Useful for domain-motif interaction studies.	Not suitable for studying PPIs related to PTMs
Assays	Affinity Determination	Uses the principle of binding affinities among proteins.	Helps finding useful interactions and analysing biological pathways.	

1.6 Molecular mimicry

Viruses are known as obligate parasites that replicate inside host cells through establishing interactions with the host proteins (Garamszegi, Franzosa et al. 2013). Basic viral infection cycle starts when a virus enters host cell, triggers host immune system and then circumvent the line of defence developed by the host cell. Adaptation processes have enabled hosts to coexist with pathogens, sometimes taking benefits from the pathogens, but most of the known pathogens are infectious leading to life-threatening human diseases (Benedict, Norris et al. 2002; Finlay and McFadden 2006). Therefore, to prevent and treat these diseases, it is crucial to understand host-pathogen biological systems (Jean Beltran, Federspiel et al. 2017). Virus-host protein-protein interactions (vhPPIs) are a regular event that occur throughout the viral life cycle. Viruses replicate through hijacking host cellular machinery i.e. transcriptional/translational machinery (Neduva and Russell 2005; Davey, Van Roey et al. 2012). The underlying mechanisms of virus interactions with host cells are still unclear and traditional molecular biology and proteomics techniques are being faced with challenges including time-consumption and cost (Chaurushiya, Lilley et al. 2012). One of the biggest challenges in terms of understanding viral diseases has been the timely discovery of viral and host proteins involved in infection cycle. Many viruses have become drug resistant which has made it even more difficult to develop successful therapeutics against them. The ideal way to eradicate viral infection is to block viral replication and this can be done by discovering host proteins and pathways being targeted by viral proteins (Garamszegi, Franzosa et al. 2013; Jean Beltran, Federspiel et al. 2017).

Studying molecular mimicry has become one of the most intriguing aspects of research. The term ‘molecular mimicry’ was first referred as sharing of antigens between pathogens and hosts (Damian 1964), also known as ‘antigenic mimicry’ which allows pathogens to hijack host cellular machinery (Kohm, Fuller et al. 2003).

1.6.1 Short linear motifs and molecular mimicry

Most of the viruses use similar strategies to mimic host motifs to control cellular pathways (Benedict, Norris et al. 2002; Finlay and McFadden 2006; Davey, Trave et al. 2011; Chaurushiya, Lilley et al. 2012). Protein interactions are often mediated by the globular domains that interact with other proteins. Globular domains and ordered regions of the proteome were once considered sole mediator of protein-protein interactions (PPIs). But recent progress in proteome research has revealed that disordered regions containing Short Linear Motifs (SLiMs) are also important mediator of PPIs. Viruses interact with the host cellular proteins through SLiMs which are like host cell SLiMs. SLiMs also known as mini-motifs and linear motifs are short stretches of amino acids (~3-10) involved in different cellular functions: post-translational modifications, PPIs, cell compartment targeting and regulation (Neduva and Russell 2005; Davey, Van Roey et al. 2012). SLiMs are known as robust and highly evolvable elements found in viruses which lead to rewiring of the vhPPIs (Chemes, de Prat-Gay et al. 2015). In most of the cases, a SLiM that is adequately exposed on protein surface can control protein stability, ligand binding and targeting, more generally, can regulate several biological pathways. Approximately, 30% of human proteome is disordered (Neduva and Russell 2005; Van Roey, Uyar et al. 2014). Most of the time, SLiMs are found in IDRs of the proteins and, sometimes in accessible loops within the folded domains, which are evolutionarily variable areas of protein where SLiMs can appear or disappear through single point mutation (Neduva and Russell 2005; Van Roey, Uyar et al. 2014). Despite the functional significance of IDRs, these disordered regions are not well studied and lack extensive characterization (Diella, Haslam et al. 2008; Hornbeck, Kornhauser et al. 2012; Nguyen Ba, Yeh et al. 2012).

SLiMs can arise *de-novo* in unrelated proteins through convergent evolution. This convergent evolution of SLiMs further complicates understanding of the interactome (Tompa and Csermely 2004). According to one estimation, there are around 1 million SLiMs in human proteome (Tompa, Davey et al. 2014), demonstrating the complexity of the regulatory mechanisms of cells. SLiMs in pathogenic proteins are known as mimicry motifs as they have similar, if not identical, amino acid composition and functions as host SLiMs. Various examples of mimicry motifs have been reported in different pathogens, especially in proteins involved in attachment, penetration and cytoadherence. One of the best-known examples in viruses is the polyproline motif (PxxPxR), which has been reported in non-structural 5A protein (NS5A) of hepatitis C virus as well as in Nef protein of HIV type 1. This polyproline motif establishes interactions with SH3 domains of the host proteins (Shelton and Harris 2008).

SLiMs are often called as molecular switches as they can switch to different functionalities with a single point mutation. This SLiM plasticity creates an Achilles heel which helps pathogenic proteins imitate host proteins and help them to interact with host cellular pathway (Davey, Trave et al. 2011). Viruses mostly interact with host proteins through establishing domain-motif interactions (DMIs) (Halehalli and Nagarajaram 2015). The current number of known DMIs in the entire human proteome is likely to be better than available stats in the database (Tompa, Davey et al. 2014). This is the reason more sophisticated methods (experimental and computational) are required to study SLiM based interactions which are important to understand the mechanism of motif mimicry in viruses. DMIs are being considered important therapeutic targets, but only few studies have been published showing the capability of DMIs as potential drug targets. Targeting these DMIs is quite challenging because of their transient, complex and promiscuous nature. Another challenging feature of DMIs is their physiochemical and structural properties (Davey, Trave et al. 2011; Corbi-Verge and Kim 2016). Understanding molecular mimicry has become an interesting area to understand how viruses hijack host cellular pathways and how viruses invade host cells. Such studies will eventually be helpful in developing novel antiviral therapeutic regimens (Dyer, Murali et al. 2007; Davey, Trave et al. 2011; Via, Uyar et al. 2015; Corbi-Verge and Kim 2016). Therefore, new motif-based strategies are much needed to study virus-host interactions (Evans, Dampier et al. 2009; Segura-Cabrera, Garcia-Perez et al. 2013).

1.7 Viral subtypes based on genetic material

During recent years, different computational methods have been developed to study proteome wide vhPPIs (Dyer, Murali et al. 2007; Evans, Dampier et al. 2009; Segura-Cabrera, Garcia-Perez et al. 2013). But most of these studies have been targeted to selected pathogens only (Emamjomeh, Goliaei et al. 2014; Barnes, Karimloo et al. 2016; Zhang, He et al. 2017). To date, there has been no study to analyse different viral subtypes based on their genetic material to see how they perturb host cellular machinery for their regulatory functions and infection cycle. Therefore, it is of interest to see how different subtypes of viruses tend to interact with host proteins through SLiMs.

1.7.1 RNA viruses

RNA viruses are considered major threat to human health and are responsible to infect millions of people around the world. RNA viruses can have single stranded RNA or double stranded RNA as their genetic material. These viruses replicate through exploiting RNA-dependent RNA polymerases. For example, retroviruses infect host cells through two copies of single stranded RNA genomes which are reverse transcribed to produce viral DNA which integrates into host DNA. RNA viruses such as Hepatitis C virus, Zika virus, Ebola virus, Yellow fever virus, Dengue virus, Polio virus, SARS, Influenza virus, retrovirus including human immunodeficiency virus and adult Human T-cell lymphotropic virus type 1) cause different human diseases (Poltronieri, Sun et al. 2015).

The RNA genome plays important role in terms of producing viral proteins necessary for viral reproduction as well as some additional duties including role as template for genomic replication, mRNA transcription, and virion assembly. In some RNA viruses, the viral genome has critical role in carrying our multiple processes in host cell. Different viral and host proteins interact with viral genome to help them achieve important functions that assist viruses to replicate in host cells. In general, protein as well as different RNA factors interact with cellular pathways to help viruses in successfully hijacking the host cell machinery (White, Enjuanes et al. 2011). Moreover, RNA viruses are known to have great structural as well as functional diversity. They can produce new RNA genome every 0.4 sec (if replication machinery is working optimally)(Moya, Elena et al. 2000). Currently, there is no effective vaccine against most of these RNA viruses therefore, it is necessary to understand how viruses infect host cells and how they replicate through hijacking host cellular machinery (Franzosa and Xia 2011).

1.7.2 Single-stranded RNA (ssRNA)

There are two main groups of RNA single stranded (ssRNA) viruses: Negative-Stranded RNA viruses (NSVs) and Positive-Stranded RNA viruses (PSVs). NSVs have single stranded RNA as their genetic material. They are further classified into two groups: segmented and non-segmented. The segmented group contains families: *Orthomyxoviridae, Arenaviridae* and *Bunyaviridae* whereas, non-segmented group has families including *Paramyxoviridae, Bornaviridae, Rhabdoviridae, Filoviridae* and *Nyamiviridae*. NSVs have highly organized genome structures in the form of nucleocapsids or ribonucleoprotein complexes where genomic RNA has association with multiple monomers of nucleoproteins (Green, Cox et al. 2014; Ortin and Martin-Benito 2015). These viruses are responsible for high mortality and morbidity rates and have caused many disease outbreaks such as influenza, measles and mumps worldwide (Ortin and Martin-Benito 2015). The life cycle of NSVs begin by attachment of the virus with the host cell where it releases its ssRNA into the cell. The released RNA is then transcribed into mRNA inside the cell and is also transcribed into a genomic strand which serves as a template to replicate the viral genome. The transcription process is carried out by the viral polymerase which is then packed inside the newly assembled virion. The replicated virion is then released outside the cell **(Figure 1.6)** (Li, Wei et al. 2013).

On the other hand, PSVs are important subgroup of RNA viruses where RNA genome is a plus stranded RNA. The RNA genome of PSVs act as blueprint for viral proteins and have cis-acting RNA elements that help regulating different viral processes including viral replication, transcription and translation (Liu, Wimmer et al. 2009; Sztuba-Solinska, Stollar et al. 2011). The life cycle of PSVs begin with the viral attachment, upon which the ssRNA is released into the host cell. The released ssRNA is then translated into a single polyprotein which is then processed into different proteins including viral polymerase and RNA-dependent RNA polymerase. A complementary strand of RNA is also generated which serves as mRNA and helps in replication process. The replicated information is then assembled as new virion and is released outside the cell **(Figure 1.6)** (Li, Wei et al. 2013).

In general, PSVs enters the host cells and replicate in the cytoplasm of the infected cells where host defence system develops unfavourable conditions for viral replication. To overcome this problem, PSVs creates intracellular environment through concentrating viral proteins which allows continuous replication of viral genome. Viral proteins hijack host factors involved in vascular trafficking and lipid biosynthesis which help protecting the viral replication machinery from host immune system, creating a safe environment for viral replication and assembly (Harak and Lohmann 2015). A known example of motif mimicry in ssRNA viruses is HCV which hijacks host cellular machinery through mimicking PxxP motifs known to bind with a variety of SH3 domains of Src kinase family (Duro, Miskei et al. 2015). Another example is the PxxP motifs in HIV which hijack host cellular machinery through establishing interactions with SH3 domain of the host (Stangler, Tran et al. 2007).

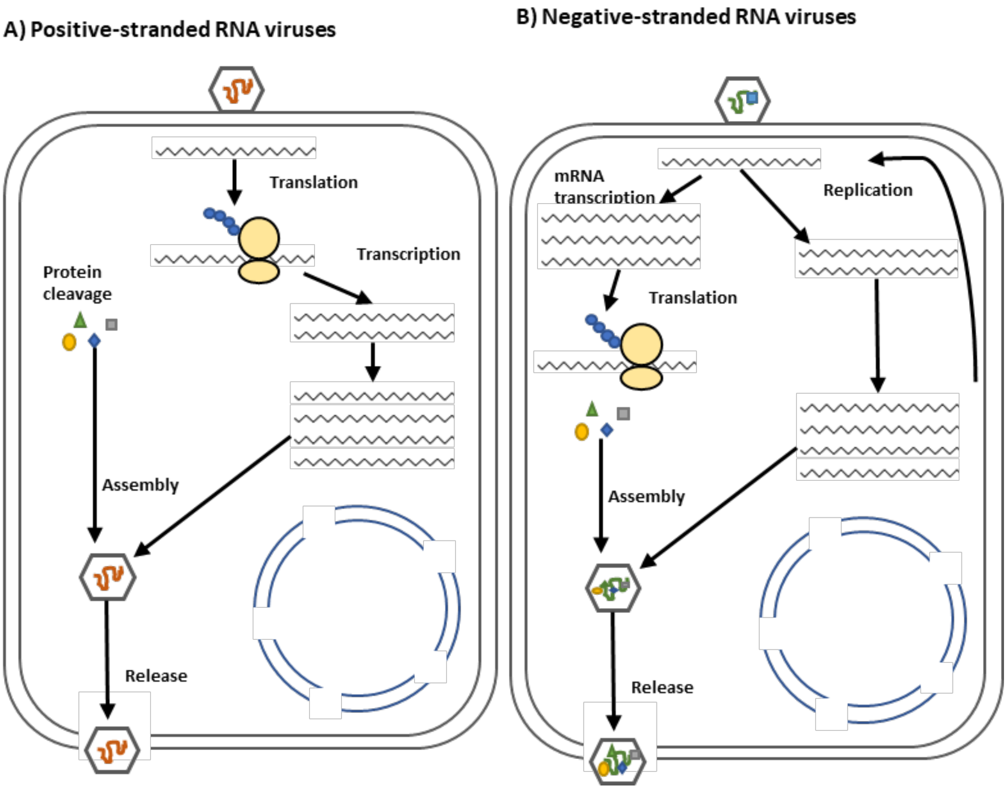


Figure 1.6. A general overview of ssRNA viral replication cycle.

A) **Replication cycle of PSVs.** Upon attachment, virus releases its plus-sense ssRNA into the host cell where it is translated to produce a single polyprotein molecule. This polyprotein is then processed into proteins i.e. RNA-dependent RNA polymerase and viral polymerase protein and a complimentary copy of RNA is produced. The new complementary RNA serves as a template for producing new plus-sense strands. The new plus-sense RNA then serves as new mRNA for replication. The replicated virus is then packaged and released from the cell.

B) **Replication cycle of NSVs.** The RNA in NSVs is transcribed into mRNA as well as can also be transcribed into full length plus-sense strand that serves as a template for replication. Viral polymerase helps in transcription which is packaged in newly assembled virion. The newly replicated virion is then released outside the cell (Li, Wei et al. 2013).

1.7.3 Double stranded RNA (dsRNA)

RNA double stranded viruses (dsRNA) are found in all types of organisms including animals, plants, fungi, terrestrial and non-terrestrial invertebrates and bacteria. The first dsRNA was discovered in reoviruses by Gamatos and Tamm in 1963 (Wickner 1993). Most of these viruses have icosahedral capsid structures and have similarity in replication strategies, biochemical and structural properties. This is the reason, cognate proteins with similar structure and functions can be identified even from distantly related viruses, providing clues on their common ancestry. dsRNA viruses are known to replicate inside host cytoplasm. These viruses invade host cells and converts ssRNA to dsRNA. Their genomic dsRNA is then transcribed into mRNA which upon translation produces proteins essential for viral replication **(Figure 1.7).** Eukaryotic systems have defence mechanisms that detect dsRNA and inactivates it through PKR or MDA5 proteins therefore, dsRNA replicate/transcribe their RNA inside icosahedral capsids (Mertens 2004).Moreover, viral proteins found in internal virion associated enzymes and innermost capsid layers are conserved in most of the viruses. One the other hand, outer capsid proteins (non-structural proteins) are diverse in their sequence as well as structural organizations (Mertens 2004). A known example where dsRNA hijacks host cellular machinery is the Segment-10 protein of Bluetongue virus which hijacks host cellular pathways through mimicking functions of different host proteins (i.e. NEDD4 and TSG101) (Wirblich, Bhattacharya et al. 2006).

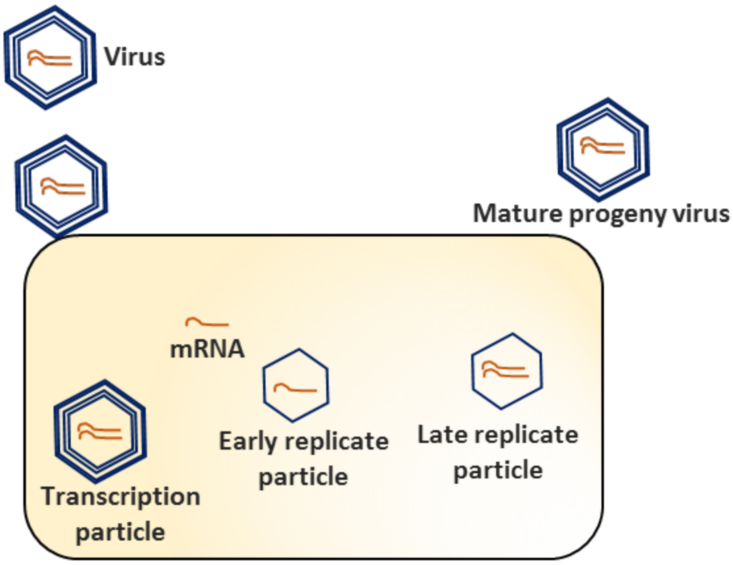


Figure 1.7. A general overview of dsRNA viral replication cycle.
Virus first attaches and enters host cell where it's dsRNA is transcribed into mRNA. The mRNA is then packaged which leads to replication of virus through producing early replicate particle. The complementary RNA is then produced leading to late replicate particle which is then released as a mature progeny virus.

1.7.4 DNA viruses

As compared to the RNA viruses, DNA viruses are less abundant and less diverse in eukaryotes. The unexpected discovery of the giant viruses (genome size bigger than bacteria, archaea and many parasitic unicellular eukaryotes) have diverted attention towards DNA viruses (Koonin, Krupovic et al. 2015).

Just like RNA viruses, DNA viruses hijack host cellular machinery to replicate their genome to increase their numbers. DNA viruses are often composed of a capsid which is capable of binding and invading host cells. Upon invasion, the virion is disassembled, and genome is released into the cell where the viral genome is transcribed into mRNA. The transcribed viral mRNA is then processed/translated into proteins (Ng, Marine et al. 2012; Krupovic and Forterre 2015). These proteins are responsible for hijacking the host cellular pathways and helps in preparing the virus to produce progeny virus. The progeny virus is then released outside the host cell and is ready to invade other cells (Ng, Marine et al. 2012; Rao and Feiss 2015). DNA viruses can be divided into two main types i.e. small DNA viruses (genome size < 10kb) and large DNA viruses (genome size >30kb). Examples of small DNA viruses include human papilloma virus (HPV) and Hepatitis B virus (HBV) while examples of large DNA viruses include Adenovirus, poxivirus and herpesvirus (Iyer, Aravind et al. 2001).

1.7.5 Single stranded DNA (ssDNA)

These are simple viruses which have single strand of DNA (ssDNA) as their genome. Most of these viruses have negative strand DNA but some of them can retain both positive and negative strands of DNA (Koonin, Krupovic et al. 2015). ssDNA viruses have evolved different invasion mechanisms depending on their hosts (bacteria, archaea and eukaryotes). For example, *Inoviridae* (filamentous bacteriophages) have evolved three diverse mechanisms of invading host cells. Some of these viruses utilize DDE transposases of IS30, IS3 or IS110/IS492 families while some encode integrases of serine/tyrosin recombinase super-families. Some *Inoviridae* and few members of *Microviridae*, hijack host XerCD recombinase machinery. On the other hand, eukaryotic ssDNA viruses integrate with host cell through endonuclease activity of their rolling circle replication initiation proteins (mimicking the mechanisms used by bacteria transposons) (Krupovic and Forterre 2015; Rosario, Mettel et al. 2018).

These viruses have one gene for encoding viral nucleocapsid and one gene for DNA encoding DNA replication enzyme. Upon invasion, these viruses need to convert their ssDNA genome into dsDNA for which they use DNA polymerase of the host cell. The dsDNA serves as a template for transcription. The transcribed RNA is then translated into viral proteins and the replicated DNA is converted back into ssDNA which is then packaged into a new virion. The new virion is then released outside the cell where it infects new cells **(Figure 1.8)** (Krupovic and Forterre 2015).

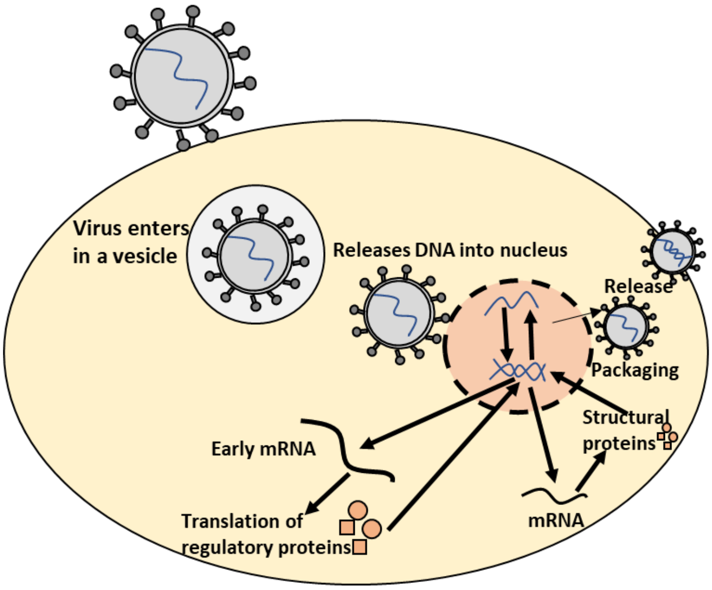


Figure 1.8. A general overview of ssDNA virus replication cycle.

ssDNA viruses start their life cycle through attachment and entry into the host cell. These viruses enter inside vesicle and upon reaching nucleus, releases their ssDNA inside it. The ssDNA is converted into dsDNA which is then transcribed into early mRNA in cytoplasm. This early mRNA is translated into different regulatory proteins which help replication of the genome. The dsDNA is then transcribed into mRNA which translates into structural proteins. The structural proteins and the newly replicated ssDNA then are packaged and released outside the cell.

1.7.6 Double stranded DNA (dsDNA)

Double stranded DNA (dsDNA) viruses have a single molecule of dsDNA as their genome (Lodish, Berk et al. 2000). Many dsDNA viral families are known to infect mammals including *Hepadnaviridae, Papillomaviridae, Polyomaviridae, Herpesviridae, Adenoviridae, Asfarviridae* and *Poxviridae*. All of these viral families except *Asfarviridae* and *Poxviridae* infect humans or animals (Koonin, Krupovic et al. 2015).

dsDNA viruses are often considered as simplest viruses to understand their life cycle. Their life cycle begins when a virus invades the host cell. Upon invasion, viral DNA enters nucleus of the host cell where it mimics the host genome. It uses host cell DNA polymerase to replicate its genome and host cell RNA polymerase for transcription. The transcribed mRNA is then transported into the cytoplasm of the host cell where it is translated into different viral proteins. Some of these proteins serve as capsid in which the newly replicated DNA is packaged. The packaged virion is then released outside of the cell where it infects other cells **(Figure 1.9)** (Kazlauskas and Venclovas 2011; Rao and Feiss 2015; Kazlauskas, Krupovic et al. 2016).

A known example of motif mimicry in dsRNA is PxLxP motif in E1 protein of Human adenovirus C serotype 5 interacting with MYND domain of BS69 protein in human. This hijacking helps viruses in regulating their viral replication during infection (Zhang, Tessier et al. 2018). Another example of motif mimicry in dsDNA viruses is where PDZ binding motif in E6 protein of human papilloma virus targets PDZ containing proteins in host cell (Accardi, Rubino et al. 2011; Segura-Cabrera, Garcia-Perez et al. 2013).

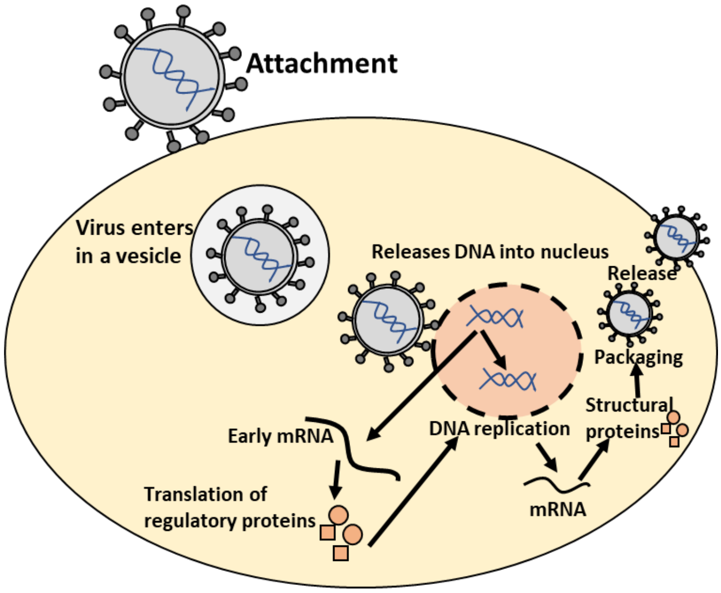

Attachment
Virus enters in a vesicle
Releases DNA into nucleus
Release
Packaging
Structural proteins
DNA replication
Early mRNA
Translation of regulatory proteins
mRNA

Figure 1.9. A general overview of dsDNA virus replication cycle.

dsDNA viruses start their life cycle through attachment and entry into the cell. They enter cell inside a vesicle and then upon reaching nucleus, releases their dsDNA molecule inside it. The dsDNA is then transcribed into mRNA in cytoplasm which translates into regulatory proteins. These regulatory proteins help in DNA replication and transcription of mRNA which then is translated into structural proteins. The newly replicated DNA and structural proteins are packaged inside a capsid and are released outside the cell.

1.8 Aims and Objectives

The main objectives were:

- Many PPI detection methods have been developed to study protein networks. However, their capability of capturing DMIs have not been evaluated comprehensively. Thus, this study was designed to see if the data being generated by high-throughput methods is actually useful for capturing SLiM mediated interactions.
- To develop algorithm to predict DMIs and evaluate enrichment of different PPI detection methods.
- To predict motif mimicry in viruses and how they hijack host cellular machinery.
- To apply computational structural biology techniques/tools to differentiate false positives and true positives.

The developed motif-based strategy can provide new insights into cellular organizations by providing clues on how two proteins interact through SLiMs and how well different methods/databases capture DMIs.

2 Chapter 2: SLiMEnrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions

2.1 Abstract

Many important cellular processes involve protein-protein interactions (PPIs) mediated by a Short Linear Motif (SLiM) in one protein interacting with a globular domain in another. Despite their significance, these domain-motif interactions (DMIs) are typically low affinity, which makes them challenging to identify by classical experimental approaches, such as affinity pulldown mass spectrometry (AP-MS) and yeast two-hybrid (Y2H). DMIs are generally underrepresented in PPI networks as a result. A number of computational methods now exist to predict SLiMs and/or DMIs from experimental interaction data but it is yet to be established how effective different PPI detection methods are for capturing these low affinity SLiM-mediated interactions. Here, we introduce a new computational pipeline (SLiMEnrich) to assess how well a given source of PPI data captures DMIs and thus, by inference, how useful that data should be for SLiM discovery. SLiMEnrich interrogates a PPI network for pairs of interacting proteins in which the first protein is known or predicted to interact with the second protein via a DMI. Permutation tests compare the number of known/predicted DMIs to the expected distribution if the two sets of proteins are randomly associated. This provides an estimate of DMI enrichment within the data and the false positive rate for individual DMIs. As a case study, we detect significant DMI enrichment in a high-throughput Y2H human PPI study. SLiMEnrich analysis supports Y2H data as a source of DMIs and highlights the high false positive rates associated with naïve DMI prediction. SLiMEnrich is available as an R Shiny app. The code is open source and available via a GNU GPL v3 license at: https://github.com/slimsuite/SLiMEnrich. A web server is available at: http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/.

Note: This chapter is published (Idrees, S., A. Perez-Bercoff and R. J. Edwards (2018). "SLiMEnrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions." PeerJ 6: e5858).

2.2 Introduction

Proteins interact with their partners through two main classes of functional modules: globular domains and Short Linear Motifs (SLiMs) (Bhattacharyya, Remenyi et al. 2006). SLiMs are short protein regions (typically 3-10 amino acids long) with a small number of key residues that mediate domain-motif interactions (DMIs) with the globular domain of a protein-protein interaction (PPI) partner (Davey, Van Roey et al. 2012). These DMIs underpin critical cellular functions, including cell cycle regulation, cell compartment targeting, post-translational modification, protein degradation, and signal transduction (Van Roey, Uyar et al. 2014). Knowledge of DMIs can provide molecular details of cellular processes and thus it is important to discover SLiMs and link them to their domain partners (Neduva and Russell 2005; Davey, Van Roey et al. 2012). Despite this, only a small fraction of the likely range of SLiMs, and the DMIs they mediate, have been identified (Tompa, Davey et al. 2014) and curated in resources such as the Eukaryotic Linear Motif (ELM) resource (Seo and Kim 2018), Linear Motif mediated Protein Interaction Database (LMPID) (Sarkar, Jana et al. 2015), interActions of moDular domAiNs (ADAN) (Kaneko, Li et al. 2008), and the database of three-dimensional interacting domains (3did) (Mosca, Ceol et al. 2014). SLiM-mediated interactions are typically low affinity (Davey, Van Roey et al. 2012) and are thus vulnerable to being overlooked by classical PPI detection methods, such as affinity pulldown mass spectrometry (AP-MS) and yeast two-hybrid (Y2H), where high stringencies are typically employed to reduce false positive interactions. Early analyses of high throughput data revealed that known SLiM-mediated interactions account for less than 1% of interactions (Neduva and Russell 2006). This was used as evidence that many more SLiMs and DMI are yet to be discovered, but also raises concerns that these methods are depleted for DMIs.

A range of computational tools now exist for the two main tasks in SLiM prediction: (1) identifying functional instances of known motifs, and (2) *de novo* prediction of new SLiM classes (Edwards and Palopoli 2015). In principle, the task of interrogating a protein sequence for known motif patterns is quite simple. Motif definitions are available from ELM (Seo and Kim 2018) and PROSITE (Hulo, Bairoch et al. 2006), and various tools exist for searching proteins for these patterns or resource-specific motif definitions (Edwards and Palopoli 2015). Other tools, like Minimotif Miner (MnM) (Lyon, Cai et al. 2018), will search sequences for similarity to known SLiMs or post-translational modifications (PTMs), but do not make motif definitions or tools available for proteome-scale searches. The short and degenerate nature of most SLiMs hampers the usefulness of predictions due to the high possibility of false positive results. This is particularly true for SLiMs with very few known occurrences, which will lack the data required for detailed modelling. It is therefore important to improve the specificity of predictions by incorporating contextual information such as evolutionary conservation and/or protein structure (Mi, Merlin et al. 2012; Krystkowiak and Davey 2017), or knowledge of interaction partners containing relevant SLiM recognition domains (*e.g.* (Kaneko, Li et al. 2008; Pichlmair, Kandasamy et al. 2012; Weatheritt, Jehl et al. 2012; de Chassey, Meyniel-Schicklin et al. 2014).

The *de novo* prediction of SLiMs is inherently more challenging and relies on assembling sets of proteins that share a SLiM. The most widespread approach is to mine PPI data to identify sets of proteins that interact with a common partner (*e.g.* (de Chassey, Navratil et al. 2008; Lieber, Elemento et al. 2010; Edwards, Davey et al. 2012). The success of prediction methods is highly dependent on the signal to noise ratio in these data, in terms of the proportion of proteins likely to contain the SLiM (Edwards, Davey et al. 2012; Edwards and Palopoli 2015). Before attempting SLiM discovery, it is therefore useful to know how well the input PPI data is capturing SLiM-mediated interactions. Different experimental parameters will influence how depleted the recovered interactions are for DMIs, and so this assessment is also useful for experimentalists when establishing an appropriate stringency threshold.

Here, we introduce a new computational pipeline (SLiMEnrich) that assesses how well PPI data are capturing DMIs and thus, by inference, how useful that data should be for SLiM discovery. The null hypothesis is that the PPI data have been generated by methods that fail to detect DMI. In this scenario, any observed DMI in the data are down to random associations between the relevant domain- and motif-containing proteins. SLiMEnrich evaluates DMI enrichment versus this null expectation through permutation tests and reports the probability of randomly recovering as many interacting domain-motif pairs as are found in the real PPI data. This enrichment evaluates datasets for the presence of DMIs, which is the prerequisite for further analysis such as SLiM prediction or calculating DMI enrichment of subnetworks versus the whole PPI network.

SLiMEnrich can use known SLiM-mediated interactions for high stringency analysis or incorporate DMI predictions by using SLiM predictions and/or known SLiM-domain interactions to expand the number of plausible DMIs in the data. Identified/predicted DMIs are returned, along with an estimated false discovery rate based on the mean number of random DMIs generated from the data. Whilst not their primary purpose, SLiMEnrich metrics can also be used to assess SLiM and/or DMI prediction strategies when applied to PPI data that is already known to contain DMIs. SLiMEnrich is therefore of potential use for both DMI prediction and assessment of PPI data. SLiMEnrich has been developed in R and implemented in Shiny to provide easy, user-friendly operation.

2.3 Materials and Methods

2.3.1 Algorithm

An overview of the SLiMEnrich pipeline is shown in **(Figure 2.1).** SLiMEnrich uses (known or predicted) SLiM occurrences, domain composition, and known SLiM interactions at the protein or domain level. These are combined to predict SLiM-mediated DMIs within pairwise PPI data supplied by the user. Input data is combined by matching protein, SLiM and Domain IDs from the input data, providing a flexible framework for analysis. PPI data is treated asymmetrically, with specified sets of putative motif- and domain-containing proteins, known as "mProteins" and "dProteins", respectively. First, SLiMEnrich identifies all possible known/predicted DMI links between mProteins and dProteins in the PPI data **(Figure 2.2).** DMI mapping can be performed using a number of different strategies depending on the desired balance of quality versus quantity of DMI **s** At one extreme, analysis can be restricted to mProtein-dProtein pairs known to interact via a DMI **(Figure 2.2, top left).** At the other extreme, mProteins with predicted SLiMs can be linked to any dProteins containing a domain known to interact with that SLiM **(Figure 2.2, bottom right).** This set of "potential DMIs" represents the overall pool of possible DMIs given the input data and mapping strategy**.**

Next, SLiMEnrich extracts "predicted DMIs" by identifying the subset of potential DMIs that are found in the PPI data, *e.g.* observed PPI pairs where the mProtein is (known or) predicted to interact with the dProtein according to the DMI strategy employed. Finally, SLiMEnrich estimates how well the PPI data is capturing DMIs by comparing the observed DMI predictions to a background distribution of expected DMIs when proteins are randomly assigned interaction partners. For this, the input PPI data is shuffled to generate 1000 random PPI datasets where each protein maintains the same number of interacting partners but the connections are randomly assigned. This is performed by first reducing PPI data to asymmetrical non-redundant protein pairs and then randomly shuffling the dProtein column whilst avoiding the introduction of redundant random PPI pairs. The random PPI datasets are then mapped onto the potential DMIs in the same fashion as the real data. Enrichment is calculated as an empirical *P*-value corresponding to the probability of seeing at least as many DMIs in random PPI data **(Figure 2.3).** A False Discovery Rate (FDR) for individual DMIs is also estimated as the proportion of the predicted DMIs explained on average by random associations, using the mean random DMI distribution capped at the observed value.

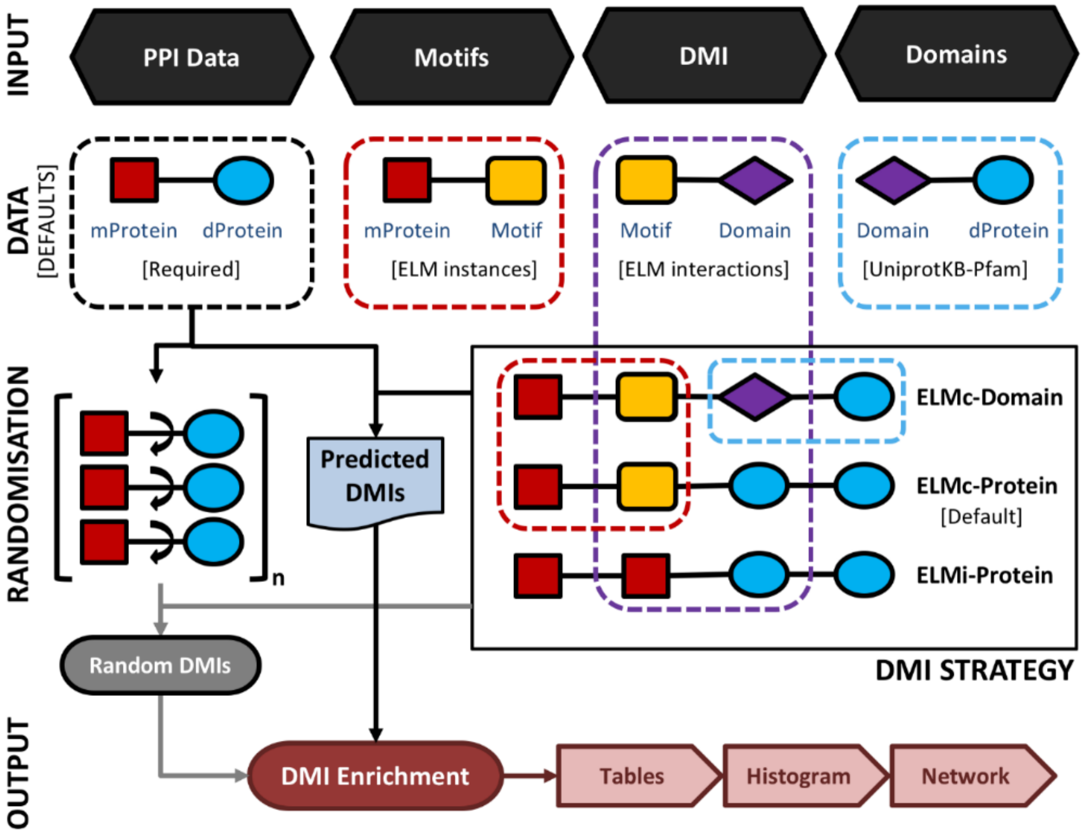


Figure 2.1. A schematic representation of the main SLiMEnrich pipeline.

SLiMEnrich takes four input files: 1. PPI data provided by the user as a set of pairwise putative motif-containing proteins ("mProteins") and their domain-containing interaction partners ("dProteins"); 2. A file providing known or predicted motif occurrences within the mProtein sequences (by default, known ELM instances are used); 3. A DMI file defining Motif-Domain interactions, relating to the DMI Strategy employed (by default, known ELM interactions are used); 4. A file that links dProteins to their domain composition (by default, human Pfam domains from UniprotKB are used). Input data is combined to establish the complete set of known/predicted “potential DMI” dependent on the DMI strategy selected (see Figure 2 and text for details): ELMi-Protein – for highest stringency, the DMI file directly links mProteins to known dProtein DMI partners (Motifs and Domains input not used); ELMc-Protein – for medium stringency, the DMI file links Motif classes to known dProtein DMI partners (Domains input not used); ELMc-Domain – for lowest stringency, the DMI file links Motif classes to known interacting Domains. Potential DMIs are then mapped on to the input PPI to identify the "Predicted DMIs" in the real data. PPI data is randomised (shuffled) 1000 times and re-mapped to potential DMIs to determine the background distribution of predicted DMIs in the case of random association (see text for details). Finally, the “Random DMI” distribution is compared to the observed “Predicted DMIs” to determine DMI enrichment in the data. Results are output in the form of a tables, a histogram of the Random DMI distribution with the observed count and empirical P-value marked, and an interactive network of the known/predicted DMIs found in the PPI data.

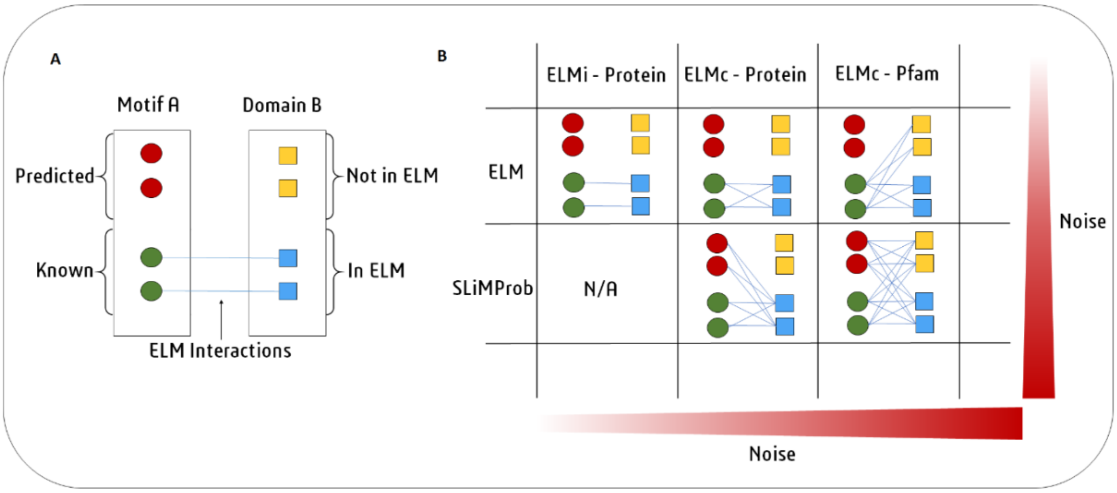


Figure 2.2. SLiMEnrich DMI prediction strategies.

SLiMEnrich uses known DMI from the ELM database to identify known DMIs or predict DMIs within the supplied PPI data. (A) In this example, Motif A is known to interact with Domain B. Motif A has two known occurrences in the data (green circles) and two predicted occurrences (red circles). Domain B is present in four proteins (squares). ELM has two annotated interactions between proteins with Motif A and proteins with Domain B (blue). (B) In the simplest and purest strategy, only known ELM interactions (ELMi) are used to assess enrichment (right panel, top left box). For small PPI datasets it might be necessary to increase the number of predicted DMI. This can be done in two ways. Top row: known motif occurrences (green circles) can be connected to all proteins known to interact with that ELM class (ELMc) (blue squares, top centre), or connected to all proteins containing a domain that interacts with that ELM class (all squares, top right). Bottom row: to increase the number of DMI further, known ELM occurrences can be replaced with SLiM predictions (all circles).

2.3.2 Requirements and Implementation

Inputs. SLiMEnrich requires a delimited pairwise PPI file as input. By default, known ELM instances (ELMi) (Seo and Kim 2018) will be used to define the motif composition of mProteins. This file can be replaced by a SLiM prediction file (generated by *e.g.* SLiMProb (Edwards and Palopoli 2015)), which has predicted SLiMs for the mProteins in the PPI file. DMIs can be predicted by one of three strategies **(Figure 2.2).** By default, the DMI file links ELM classes (ELMc) directly to dProteins using known ELM binding partners (Seo and Kim 2018). For more stringent analysis, these binding partners can be linked directly to specific ELM-containing proteins, in which case the DMI file links mProteins and dProteins, and the motif occurrence file is ignored **(Figure 2.1).** For more relaxed/flexible analysis, the DMI file will link motifs to binding domains, which are then linked to dProteins via a domain composition file. By default, SLiMEnrich uses Pfam domains (Finn, Coggill et al. 2016) for reviewed human Uniprot proteins (The UniProt 2017) and links them to ELM-binding domains (Seo and Kim 2018). If alternative data sources are used, users should also provide a file of protein-domain links for the dProteins in the PPI file, and/or a motif-domain file that defines the known domain-motif interactions. Note that this can be used to interrogate PPI data for enrichment of any interaction type. For example, two protein-domain files could be linked through known domain-domain interactions. Alternatively, the ELMi-Protein DMI strategy enables the enrichment analysis of any set of PPIs, allowing SLiMEnrich to examine overlaps between PPI datasets. Default fields for user files ("mProtein", "dProtein", "Motif", "Domain") are shown in **Figure 2.1**, and can be set to custom values in the SLiMEnrich App.

Example data. SLiMEnrich comes with example data of Adenoviridae proteins and their human interactors downloaded from the PHISTO database (2017-07-26) (Durmus Tekir, Cakir et al. 2013). ELM (downloaded 2018-07-17) (Dinkel, Van Roey et al. 2016) regular expression matches in the viral proteins were predicted using SLiMProb v2.5.0 (Edwards and Palopoli 2015) with disorder masking. A table of ELM-binding Pfam domains was downloaded from ELM (2018-07-17) (Dinkel, Van Roey et al. 2016). Pfam domains for human proteins were extracted from Uniprot (downloaded 2017-03-08) (The UniProt 2017).

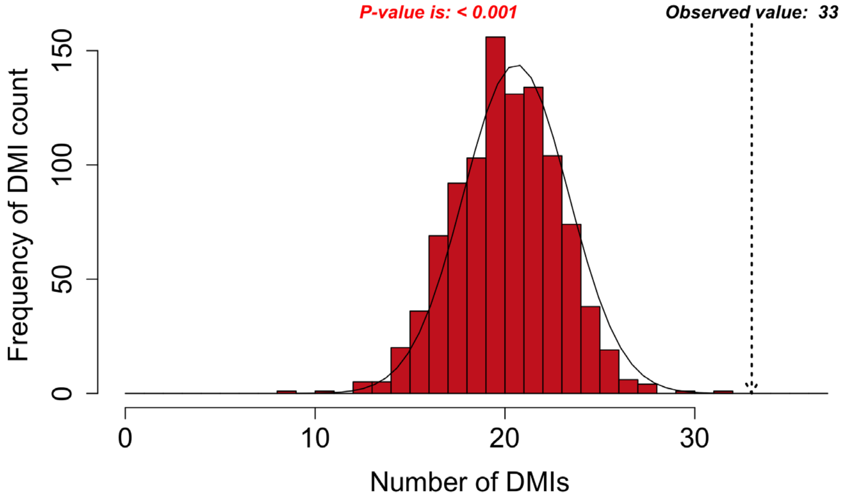


Figure 2.3. DMI enrichment histogram for SLiMEnrich example data.

Histogram of DMI enrichment in example data for Adenoviridae proteins and their human interactors (see text for details) from the SLiMEnrich app, using the most permissive ELMc-Domain DMI strategy and SLiMProb motif predictions. Frequency bars indicate the number of randomised PPI datasets returning a given number of predicted DMIs. The dotted arrow indicates the observed number of predicted DMIs in the real data.

Outputs. The primary output of SLiMEnrich is the observed number of known/predicted DMIs compared to the distribution from the randomised PPI data **(Figure 2.3).** SLiMEnrich also provides tables of both "potential DMIs" and "predicted DMIs" **(Figure 2.1, see Algorithm for details)**, summary plots of predicted DMI numbers and an interactive DMI network **(Figure 2.4)**. Together, these enable the user to explore the data for proteins, SLiMs and/or domains that might be biasing results. This can be seen with the example Adenoviridae analysis, where the Pkinase domain (PF00069) mediates a large proportion of the predicted DMIs via multiple modification ELMs **(Figure 2.4),** which will inflate the probability of DMIs in the random PPI data. Tables can be downloaded as comma-separated text files. The summary plots, enrichment histogram and DMI network can be downloaded as PNG files.

Implementation. SLiMEnrich is a standalone application written entirely in R. It is platform independent and can be launched locally from any R environment (e.g. RStudio). SLiMEnrich takes advantage of the reactive programming feature of Shiny to cache computational steps to avoid unnecessary computing during an interactive session. The code is open source and available via a GNU GPL v3 license at: https://github.com/slimsuite/SLiMEnrich. SLiMEnrich is also implemented as a Shiny webserver at: http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/. Additional details can be found at: https://github.com/slimsuite/SLiMEnrich/wiki.

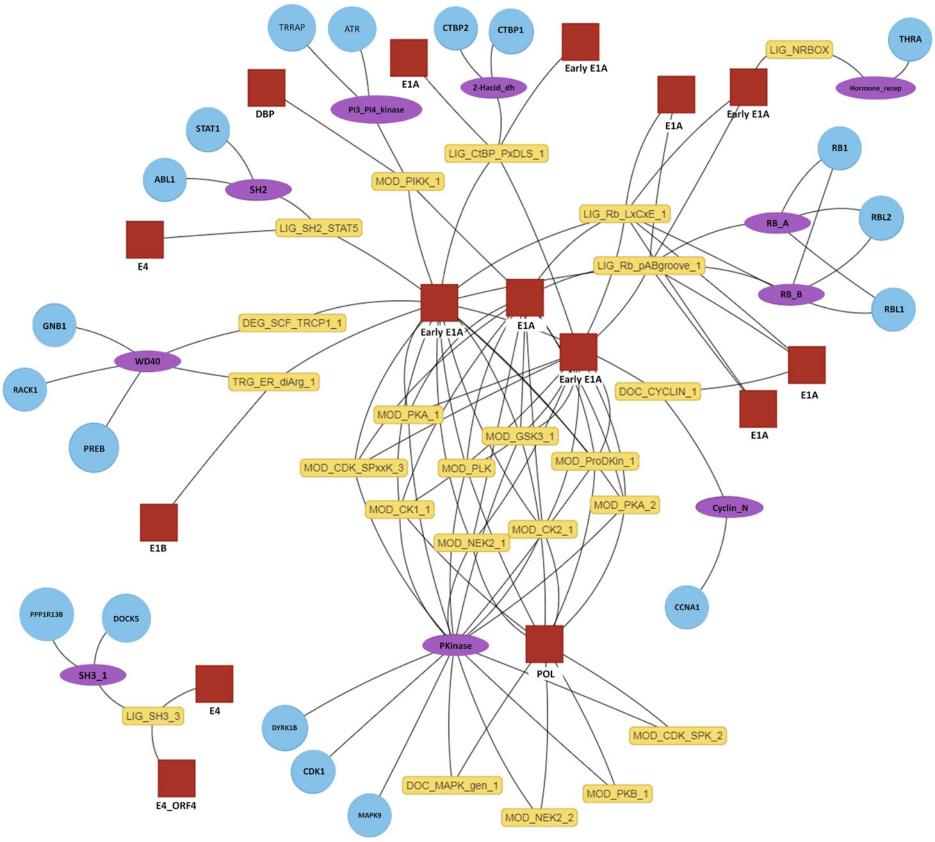


Figure 2.4. Interactive predicted DMI network for example data.

Predicted DMIs for example Adenoviridae proteins and their human interactors, using the most relaxed strategy (predicted SLiMs connected via domains, see text for details). Several layout options are provided and nodes can be manually positioned. The protein, domain and motif identifiers used in the network are determined by the user input. Using default data, these will be UniprotKB, Pfam and ELM identifiers. For this example, UniprotKB identifiers have been mapped onto HGNC gene symbols and Pfam identifiers onto Pfam domain names. Red square, motif-containing protein ("mProtein"); Yellow box, motif; Purple ellipse, domain; Blue circle, domain-containing protein ("dProtein").

2.3.3 Case study: Domain-motif resolved yeast-two-hybrid human interactome

Pairwise human PPIs were extracted from a high-throughput human Y2H study that detected ~14,000 binary interactions (Rolland, Tasan et al. 2014) and converted into a non-redundant, symmetrical PPI dataset of 26,166 mProtein-dProtein PPIs (*i.e.* with each PPI pair present as P1-P2 *and* P2-P1), restricted to reviewed Uniprot proteins. Protein sequences were downloaded from Uniprot (2017-03-01). A list of ELMs and their domain partners was retrieved from the ELM database (2018-07-17) (Dinkel, Van Roey et al. 2016). ELM occurrences in the human proteins were predicted by SLiMProb v2.5.0 (Edwards and Palopoli 2015) with disorder masking (IUPred (Dosztanyi, Csizmok et al. 2005), cut-off 0.2 (Edwards, Davey et al. 2007)) to restrict analysis to low stringency predicted disordered regions. Pfam domains were parsed from Uniprot entries using SLiMBench (Palopoli, Lythgow et al. 2015). Splice isoforms for all data were mapped onto their parent Uniprot identifier. SLiMEnrich was used to map known and predicted DMIs onto the Y2H dataset using five strategies of decreasing stringency: (1) known ELM PPIs only, (2) known ELM instances mapped onto proteins known to interact with the ELM class, (3) known ELM instances mapped onto Pfam domains known to interact with the ELM class, (4) SLiMProb predictions mapped onto proteins known to interact with the ELM class, (5) SLiMProb predictions mapped onto Pfam domains known to interact with the ELM class **(Figure 2.2).**

2.4 Simulation of poor-quality SLiM predictions

SLiMEnrich is not a DMI prediction tool *per se* and should not require completely accurate SLiM occurrence data to identify enrichment indicative of PPI data that captures DMIs. To investigate the impact of noisy SLiM prediction data, we replaced increasing proportions (25%, 50%, 75% and 100%) of the known ELM instances (2018-07-17) (Dinkel, Van Roey et al. 2016) with random occurrences and repeated analysis of the Y2H interactome case study. This was performed by replacing different proportions of the ELM proteins (*i.e.* proteins containing a known ELM) with a protein randomly selected from reviewed human Uniprot proteins (The UniProt 2017).

For direct comparison, the distribution of normalised predicted DMIs, D, was calculated as follows:

$$D = \frac{O - R}{\bar{R}},$$

where O is the observed predicted DMI count, R is the distribution of random predicted DMIs, and $\bar{R}$ is the mean random predicted DMI count.

2.5 Results

2.5.1 Case study: Domain-motif resolved yeast-two-hybrid human interactome

SLiMEnrich analysis revealed the case study Y2H data to be enriched for DMIs under all DMI prediction strategies **(Table 2.1).** Restricting analysis to known DMIs identified fourteen in the Y2H data, which represented a more than 100-fold enrichment over the random expectation (mean 0.122). Including DMIs where a dProtein was known to interact with the ELM class **(Figure 2.2, centre column)**, almost doubled the number of predicted DMIs but with nearer six times more random DMIs on average, reducing the enrichment over three-fold. Including DMIs where a dProtein contained Pfam domain known to interact with the ELM class **(Figure 2.2, right column)** dramatically increased the numbers of both predicted and random DMIs, with a corresponding drop in enrichment. Using SLiMProb predictions in place of known ELMs **(Figure 2.2, bottom row)** similarly increased both predicted and random DMIs, decreasing enrichment. In all cases, none of the 1000 randomised datasets matched or exceeded the observed number of predicted DMI, making the enrichment strongly significant (P < 0.001).

Table 2.1. SLiMEnrich analysis of Y2H case study using different DMI prediction strategies.

Strategy	Known: ELMi-Protein	Known: ELMc-Protein	Known: ELMc-Domain	SLiMProb: ELMc-Protein	SLiMProb: ELMc-Domain
Potential DMI (NR)	62	164	6,314	39,572	969,380
Predicted DMI (NR)	14†	25	74	204	1,524
Mean Random DMI (3 s.f.)	0.122	0.830	9.76	139	1,310
p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Enrichment (3 s.f.)	115	30.1	7.58	1.47	1.16
FDR (4 d.p.)	0.0087	0.0332	0.1319	0.6820	0.8602
Unique mProteins*	13	22	52	175	768
Unique ELM classes*	N/A	16	40	35	128
Unique Pfam domains*	N/A	N/A	30	N/A	51
Unique dProteins*	10	17	53	36	366

* Unique counts correspond to Predicted DMI.

† Known DMI from ELM database.

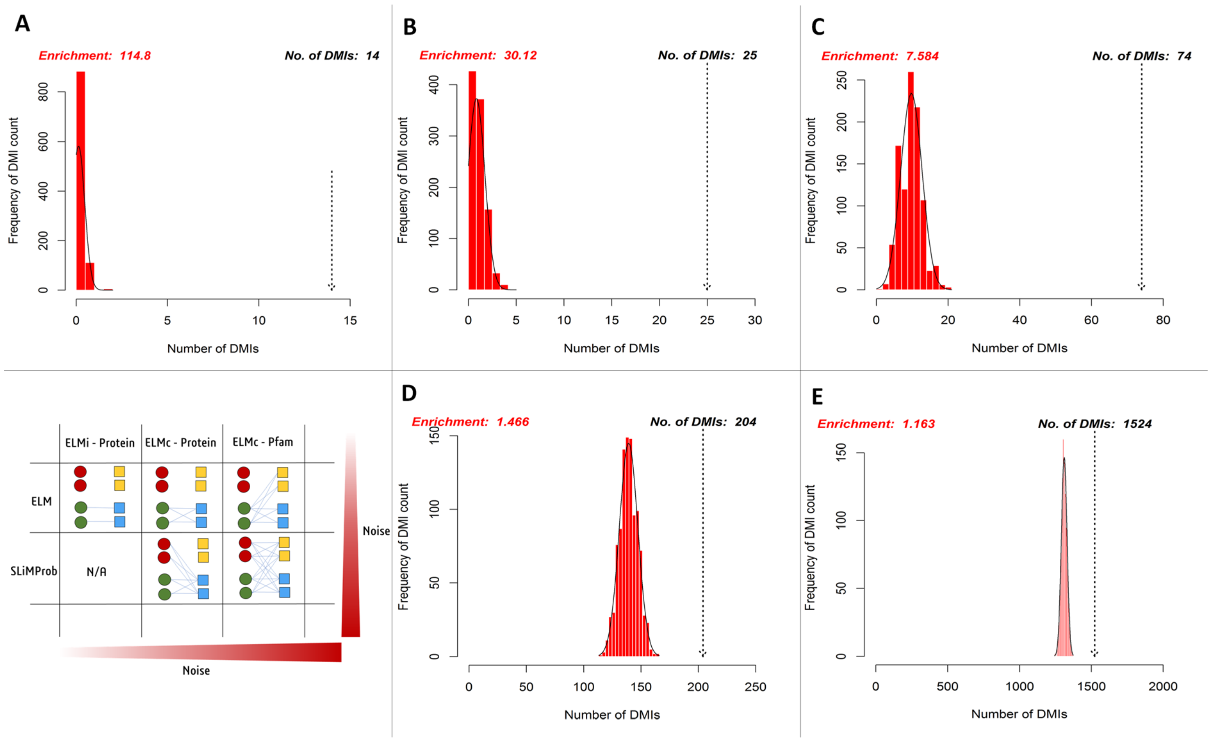


Figure 2.5. Enrichment statistics and histogram of expected random DMI counts in human Y2H case study data using known and predicted ELM instances.

Frequency bars indicate the number of randomised PPI datasets returning a given number of predicted DMIs. The dotted arrow indicates the observed number of known or predicted DMIs in the Y2H data (see text for details). DMI prediction strategies match those in Figure 2.2 (see text for details): (A) known ELM occurrences connected to interacting proteins; (B) known ELM occurrences mapped to proteins known to interact with that motif class; (C) known ELM occurrences mapped to proteins containing a domain known to interact with that motif class; (D) SLiM predictions mapped to proteins known to interact with that motif class; (E) SLiM predictions mapped to proteins containing a domain known to interact with that motif class.

2.6 Simulation of poor-quality SLiM predictions

To directly compare the effects of replacing real ELM-containing proteins with random human proteins in different proportions (25%, 50%, 75%, 100%), the distribution of normalised predicted DMI, *D*, in the Y2H data was compared for each dataset **(Figure 2.6).** *D* is the distribution of expected true positive predicted DMIs, normalised to units of mean random predicted DMIs, *i.e.* *D* = 1 is equivalent to FDR = 50%; enrichment is 1 + mean *D*. The more permissive domain-based DMI prediction strategy **(Figure 2.2, top right)** was used, as the numbers of predicted DMIs for more stringent strategies were very small **(Table 2.1)** and this strategy still showed strong (7.6x) DMI enrichment in the data **(Figure 2.5).** Despite the decline in enrichment scores with increasing proportions of random motif occurrences, enrichment remained significant even when 75% of the real data was replaced **(Figure 2.6).**

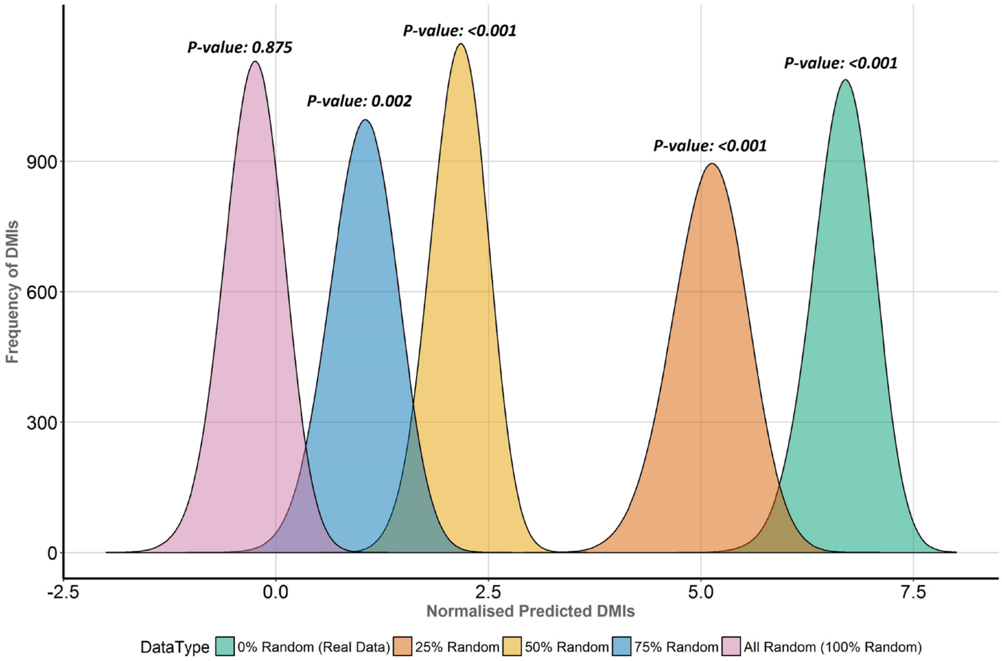


Figure 2.6. Enrichment analysis of known DMIs in human Y2H case study data with increasing proportions of random motif instances.

SLiMEnrich results for known ELMs in the human Y2H case study data mapped using the ELMc-Domain strategy, converted into the normalised number of predicted real DMIs (see text for details). Higher normalised predicted DMI counts indicate greater DMI enrichment, with zero marking no enrichment over random. Green (furthest right) is the real data using all known true positive ELM instances. The other curves (right to left) represent distributions for four randomised datasets where increasing proportions (25%, 50%, 75% and 100%) of ELM proteins were replaced with random human proteins.

2.7 Discussion

Using PPI data for SLiM discovery faces something of a contradiction. Due to their scale, data from high throughput PPI detection studies are where the novel interactions are most likely to be found. However, high stringency filters are often applied to high throughput methods to increase confidence in individual interactions, with the concomitant concern that low affinity DMIs will be lost as a consequence. The primary purpose of SLiMEnrich is to address this concern by assessing how well a given PPI dataset is capturing DMIs. Where PPI datasets are large, this assessment can be restricted to a high-quality set of known DMIs. Where the number of known DMIs in the data becomes prohibitively small, predicted DMIs can supplement or replace the known DMIs.

A detailed analysis of different PPI data sources is the subject of future study and beyond the scope of this paper. Here, we present a case study to illustrate the use of SLiMEnrich to analyse the DMI enrichment in a single PPI dataset. We have applied five different DMI identification/prediction strategies **(Figure 2.2)** to a high-throughput Y2H human PPI study (Rolland, Tasan et al. 2014) **(Table 2.1, Figure 2.5)**. On face value, the ability of the Y2H PPI data to capture known DMIs might be considered disappointing. Only 14 of the 590-known human DMI protein pairs in ELM (2.37%) were found in the 26,166 PPI considered. This is consistent with earlier analyses that have highlighted the rarity of known SLiM-mediated interactions in high throughput PPI data (Neduva and Russell 2006). However, even this modest numbers reflects a massive enrichment (approx. 115-fold) over the expected number of known DMIs to occur in the PPI data by chance. Whilst we cannot rule out unexpected confounding factors, such as additional high affinity interactions between pairs of proteins that also share a DMI, this implies that the low absolute numbers are due to the small number of known DMIs rather than the inability of Y2H methods to detect DMIs. Considered analysis has estimated that the human proteome has in the order of 100,000 SLiMs involved in DMIs (ignoring post-translational modifications) (Tompa, Davey et al. 2014), which is orders of magnitude greater than the known DMIs in ELM (Seo and Kim 2018). Overall, SLiMEnrich results indicate that these data are indeed capturing real SLiM-mediated interactions and are therefore suitable for *de novo* SLiM prediction. This, in turn, increases confidence in previous large scale SLiM predictions (de Chassey, Navratil et al. 2008; Lieber, Elemento et al. 2010; Edwards, Davey et al. 2012); these often rely on rediscovery of known motifs as validation, which could be biased by incorporation of literature-based high confidence DMIs in the PPI data.

Employing a less stringent DMI identification strategy predictably boosted the numbers of predicted DMIs and continued to reveal significant enrichment in the Y2H data despite the possible incorporation of possible false positive SLiM and/or DMI predictions **(Table 2.1, Figure 2.5).** As expected, the enrichment decreased as the noisiness of the data increased, although the enrichment remained highly significant. This was supported by analysis where real ELM-containing proteins were replaced with random human proteins to simulate noise **(Figure 2.6).** Taken together, these results indicate a degree of robustness of the SLiMEnrich approach to the quality of the SLiM data. However, they also highlight a lack of robustness in the individual DMI predictions. For the purest known DMI analysis (linking known ELM instances to known ELM-interacting *proteins*), most randomised datasets did not return a single DMI. It is therefore highly likely that the 11 additional DMIs discovered by the ELMc-Protein strategy are real DMIs. The cost is that the low numbers might affect the accuracy with which the mean random DMI count, and thus enrichment, can be calculated. Relaxing the strategy to use SLiMProb predictions and/or allow DMI predictions based on interactions between ELM classes and Pfam domain classes, substantially increased the numbers of predicted DMIs but dramatically reduced the observed enrichment for both known and predicted SLiM occurrences. Using predicted SLiMs, it should be noted that the estimated false positive rate for individual DMI predictions is very high (*FDR*=0.86 when linking predicted SLiMs via ELM-binding Pfam domains). This highlights the need for caution when interpreting naïve large-scale predictions of this nature. As illustrated for the Adenoviridae-human PPI example data **(Figure 2.4),** random numbers for the Y2H case study will be inflated by a large over-prediction of kinase domain-mediated DMIs, as well as other domains with a specificity of interaction not captured at the level of Pfam definitions. Users may wish to screen out promiscuous domains and/or motifs if low stringency approaches are required to get sufficient DMI numbers.

2.7.1 Using SLiMEnrich to assess enrichment of different PPI types

Although the focus of SLiMEnrich is on DMIs, the approach is flexible and can be easily adapted to other PPI types. Direct analogues of DMIs can be studied by replacing the motifs with a different interaction feature, *e.g.* replacing motifs with domains to investigate enrichment of DDIs. More simply, SLiMEnrich could be used to study the overlap between two different PPI datasets, accounting for the connectedness of the proteins involved, by replacing the known ELM interactions with any source of pairwise PPIs. Although the PPI data for the case study was made symmetrical, the asymmetrical handling of the PPI data by SLiMEnrich would even allow intra-dataset comparisons, such as examining the overlap between PPIs when proteins are baits versus preys in a Y2H or pulldown experiment.

2.7.2 Using SLiMEnrich to assess DMI predictions

Once it has been established that a given PPI dataset, such as the Y2H case study presented, is enriched for DMIs versus the random expectation, it is also possible to use SLiMEnrich on these data to compare different DMI predictions and prediction strategies **(Figure 2.5, Figure 2.6, Table 2.1).** Several studies have combined PPI data with computational approaches to identify new DMIs for known recognition domains, such as SH2, SH3, PDZ and WW domains (e.g. Encinar et al. 2009; Kelil et al. 2016; Luck et al. 2011; Weatheritt et al. 2012). SLiMEnrich could be used to assess the source PPI data, or to compare/validate the resulting predictions on independent unbiased PPI data. Targeted experimental methods have also been developed and applied to find motif mediated interactions. Again, most of these methods have been applied to a limited set of domain families such as PDZ, SH2, SH3, and WW which has left many important domain families out of the picture

(Blikstad & Ivarsson 2015). SLiMEnrich could be used to verify that these methods are successfully targeting DMIs, or use them as experimental PPI data for validating in silico DMI predictions. Similarly, SLiMEnrich could use quality PPI data to assess the enrichment of novel SLiM-mediated interactions, predicted by tools such as PepSite (Petsalaki et al. 2009; Trabuco et al. 2012) and PIPER-FlexPepDock (Alam et al. 2017; Kozakov et al. 2006), which predict protein-peptide binding from modelling of three dimensional structures. Network-based approaches have been applied at large-scale, combining predicted motif instances with PPIs and protein domain composition (e.g. Garamszegi et al. 2013; Kim et al. 2014). Over-prediction of SLiMs can translate into over-prediction of DMIs when identifying PPI pairs where one protein has a predicted SLiM that is known to interact with a globular domain found in the other protein (Garamszegi et al. 2013; Horn et al. 2014; Weatheritt et al. 2012). This will be particularly true for PPI data in which DMIs are poorly represented, either because DMIs are not efficiently captured, or because of abundant false positive interactions (von Mering et al. 2002), which will increase the proportion of spurious protein-motif-domain-protein interaction linkages. Combining PPI data from many sources is attractive but runs the risk of generating false enrichments through the inclusion of data from low throughput focused studies. SLiMEnrich provides a useful mechanism for estimating both the enrichment and the false discovery rate of such predictions on different PPI sources; as highlighted for our case study, significant enrichment may still have a high false discovery rate.

2.8 Conclusion

There are many data- and method-specific factors that will determine whether protein-protein interaction (PPI) data are useful for short linear motif (SLiM) prediction. The presence of real domain-motif interactions (DMIs) is a baseline requirement that is generally assumed but rarely tested. SLiMEnrich is an open source R application that will identify known or predicted DMIs in PPI data and estimate how well that PPI data is capturing DMIs compared to randomised PPIs. This estimate is useful for identifying suitable PPI data for *de novo* SLiM prediction. SLiMEnrich statistics also estimate the confidence in individual DMI predictions, enabling assessment of methods that aim to improve the specificity of DMI predictions by filtering SLiM predictions and/or PPI data. Users can run SLiMEnrich online (http://shiny.slimsuite.unsw.edu.au/SLiMEnrich/) or download the code for local use (https://github.com/slimsuite/SLiMEnrich).

3 Chapter 3: High-throughput PPI data as a source of capturing domain-motif interactions (DMIs) and domain-domain interactions (DDIs)

3.1 Abstract

Protein-proteins interactions (PPIs) are vital in carrying out different cellular functions. There are many data- and method-specific factors that determine whether PPI data is useful for domain-motif interactions (DMI) or domain-domain interactions (DDI) prediction. The presence of real DMI/DDI is a baseline requirement that is generally assumed but rarely tested. To see which PPI detection method could be better at DMI/DDI prediction, we conducted a comparative study of currently available proteome wide human interactomes. Different publicly available datasets of leading high-throughput methods (Y2H, AP-MS and CoFrac-MS) were compared to see their capability of capturing DMIs and DDIs. SLiMEnrich was employed to evaluate enrichment of DMIs and DDIs in different PPI datasets using known DMI/DDI information. It was found that high throughput methods were not notably worse than PPI databases and, in some cases, seem a lot better. BioPlex2.0 and HI-II-14 were the best scorer in terms of capturing DMIs and DDIs whereas, CoFrac interaction data wasn't found to be as good for capturing DMI/DDI. Comparison of Y2H and AP-MS interactions available in three well known PPI databases i.e. BioGrid, IntAct and HIPPIE revealed that both methods were good at predicting DMIs as well as DDIs. Overall, it can be concluded that all PPI datasets were indeed capturing DMIs and DDIs with significant enrichment (P-value < 0.001) and both Y2H and AP-MS can be a reliable method to predict DMIs and DDIs.

3.2 Background

During the past decade, different studies have been conducted to discover PPIs in different organisms (De Las Rivas and Fontanillo 2012; Blikstad and Ivarsson 2015; Lum and Cristea 2016; Luck, Sheynkman et al. 2017; Peng, Wang et al. 2017). The knowledge generated by these studies is being widely used to get insights into the cellular organization of the organisms as well as to cure different diseases including cancer, viral and bacterial infections through targeting PPIs and disrupting signalling events (Lubovac, Gamalielsson et al. 2006; Seo and Kim 2018). The availability of reference proteome maps and improvements in PPI detection assays are being considered important in terms of mapping large proportion of the human interactome (Kim, Pinto et al. 2014; Wilhelm, Schlegl et al. 2014).

SLiMs interact with domains of other proteins to establish domain-motif interactions (DMIs) which are often transient and of low affinity (1–150 μM range) (Tompa and Csermely 2004; Diella, Haslam et al. 2008; Dinkel, Van Roey et al. 2016). Despite the significance of DMIs in mediating important cellular functions, the current knowledge of DMIs is still lacking and it can be said that the current number of known DMIs in resources like ELM (Dinkel, Van Roey et al. 2016) and 3DID (Mosca, Ceol et al. 2014) is likely to be better than available stats (Davey, Van Roey et al. 2012; Bhowmick, Guharoy et al. 2015; Peng, Wang et al. 2017; Seo and Kim 2018).

3.2.1 High-throughput methods and current challenges

During the past few years, different experimental techniques have been developed to detect PPIs. Each PPI detection technique has their own advantages and disadvantages. Small scale experiments can detect low number of PPIs, but their quality of interactions is often high. On the other hand, high-throughput methods can detect large number of PPIs, but the interaction quality is often low (higher rate of false positives and false negatives) (Lum and Cristea 2016; Luck, Sheynkman et al. 2017). The three well-known high-throughput methods to detect large number of PPIs are Affinity Purification coupled Mass Spectrometry (AP-MS), Yeast two hybrid (Y2H) and Co-fractionation coupled Mass Spectrometry (CoFrac-MS) (Yu, Braun et al. 2008).

Y2H is being considered as a powerful tool to discover direct binary interactions between proteins **(Figure 1.3).** The problem with this technique is its protocol specific biasness (i.e. condition-specific, transient and inter-complex interactions) (Yu, Braun et al. 2008) and higher false discovery rate (i.e. false positives and false negatives). Thus, Y2H is not considered as a good technique to obtain precise binary map within complexes (Deane, Salwinski et al. 2002; Kuchaiev, Rasajski et al. 2009).

AP-MS on the other hand is considered good to detect co-complex interactions which includes both direct and indirect interactions **(Figure 1.4).** The main issue with this technique is to distinguish between direct and indirect interactions (Teng, Zhao et al. 2015). Similarly, CoFrac-MS has been found to be more suitable for identifying co complex interactions, including direct and indirect interactions between proteins **(Figure 1.5)** (Kim, Sabharwal et al. 2010). Just like AP-MS, this method can be used at proteome level, but it's often difficult to distinguish between direct or indirect interactions between protein pairs (Luck, Sheynkman et al. 2017).

The data generated by these PPI detection methods is being used to study biological pathways, protein complexes, protein functionality and to identify potential drug targets. However, to ensure that the knowledge gained by studying PPIs is biologically meaningful, it is important to ensure the quality of the detected PPIs (Stein and Aloy 2008; Kim, Sabharwal et al. 2010). These high-throughput methods are likely to capture false positive interactions therefore, experimental as well as computational methods to validate these interactions are much needed (Stein and Aloy 2008; Kim, Sabharwal et al. 2010). Another major problem often faced to detect PPIs is their physiological settings during the experiment as certain PPIs occur at certain conditions. Moreover, several factors can also influence PPI detection including transient nature of interactions, PTMs, abundance of proteins and IDRs. Thus, unraveling proteome wide interactome is quite challenging. The two most common ways to evaluate reliability of PPIs is to design new experimental methods to validate the interactions or to develop new computational methods to find likely to be true positive interactions through filtering out possible false positives from the data by finding probability of the observed PPI (Pitre, Alamgir et al. 2008; Kim, Sabharwal et al. 2010; Idrees, Perez-Bercoff et al. 2018).

The data generated by current high-throughput screens is being considered important in terms of discovering novel SLiMs and DMIs, however, due to high stringency filtering, low affinity DMI are likely to be lost (Stein and Aloy 2008; Kim, Sabharwal et al. 2010). This in-fact raises the concerns that current high-throughput methods might be depleted for DMI and it's important to utilize appropriate PPI detection techniques when discovering SLiMs or DMIs (Neduva and Russell 2006).

Previously, we have developed an algorithm called SLiMEnrich (Idrees, Perez-Bercoff et al. 2018) which is available as an online application as well as a command-line program to assess the probability of observed DMIs in a given dataset (see Chapter 2 for details). We looked at a single Y2H dataset and observed significant enrichment and therefore, decided to implement it to other large scale human interactomes. This chapter is focused on the implementation of SLiMEnrich to evaluate different high-throughput methods as well as databases to see which one is better at capturing different sorts of interactions (i.e. DMI and DDI). The outcome of this study will give a better understanding of how different methods capture different sorts of interactions.

3.2.2 Aims and Objectives

To date, there has been no specific study to validate the efficiency of high-throughput methods as a source of capturing different sorts of interactions (i.e. DMI and DDI) (Blikstad and Ivarsson 2015; Seo and Kim 2018) therefore, we have conducted a proteome wide comparative study to validate different high-throughput methods and databases to help systems biologist choose appropriate methods when discovering SLiMs, DMIs or DDIs.

The main objectives of this study were:

- To see if data being generated by high-throughput methods is useful for capturing DMIs and DDIs.
- To evaluate public databases as a source of capturing DMIs and DDIs.
- To see if any particular method or database is better at capturing DMIs or DDIs.
- To evaluate performance of binary vs co-complex high-throughput methods as a source of capturing DMIs and DDIs.

3.3 Methods

3.3.1 Data collection and processing

SLiM data was downloaded from the Eukaryotic Linear Motif (ELM) database (Dinkel, Van Roey et al. 2016), which contains manually curated and experimentally validated SLiM data from the literature. This makes ELM a highly reliable SLiM resource. 245 ELM classes (e.g. distinct SLiMs) with experimentally validated motif instances (2896 specific protein occurrences), associated interacting domain data (308 ELM interacting domains) and known human DMIs (789 human DMIs) were downloaded from [http://www.elm.eu.org/] on 2018-07-17. Five publicly available high-throughput interaction datasets were downloaded: HI-II-14 (Rolland, Tasan et al. 2014), CoFrac-12 (Havugimana, Hart et al. 2012), CoFrac-15 (Wan, Borgeson et al. 2015), BioPlex2.0 (Huttlin, Bruckner et al. 2017) and QUBIC-15 (Hein, Hubner et al. 2015). In addition, five well-known PPI databases were evaluated: two comprehensive databases, BioGrid (Oughtred, Chatr-Aryamontri et al. 2016) and IntAct (Hermjakob, Montecchi-Palazzi et al. 2004); two high quality human PPI databases, the High quality INTeractome (HINT) database (Das and Yu 2012) and the Human Protein Reference Database (HPRD) (Keshava Prasad, Goel et al. 2009); and one meta-database that integrates data from multiple PPI databases, the Human Integrated Protein-Protein Interaction rEference (HIPPIE) (Alanis-Lobato, Andrade-Navarro et al. 2017). IntAct and BioGrid were reduced to human interactions only. Datasets were first mapped onto Uniprot IDs, were restricted to pairs of reviewed Uniprot proteins only and were treated as non-redundant symmetrical interactions. A False Discovery Rate (FDR) for individual DMIs is also estimated as the proportion of the predicted DMIs explained on average by random associations, using the mean random DMI count.

3.3.1.1 PPI subsets by experiment type

PPI subsets by experiment type (AP-MS, Y2H and Co-fractionation) were made for BioGrid, IntAct and HIPPIE. Keywords for pulling interactions from BioGrid database were "Two-hybrid", "Co-fractionation" and "Affinity Capture-MS". BioGrid had two-hybrid interactions based on low as well as high-throughput screens. Only high-throughput two-hybrid interactions were selected. For the IntAct database, molecular interaction ontologies were used to pull subsets: MI:0676, MI:0400 and MI:0004 for pulling affinity purification interactions and MI:0018 for pulling two-hybrid interactions. "Affinity", "two-hybrid" and "co-fractionation" were used as keywords to extract interactions from HIPPIE database. All PPI datasets were restricted to reviewed Uniprot proteins, were made symmetrical and redundancy was removed. The analysis here is focused on directed network having specific motif and domain proteins therefore, it is essential that both A-B and B-A are in the analysis. The percentage (%) of PPI explained was calculated using non-redundant symmetrical PPI pairs (i.e. A-B also explains B-A).

3.3.2 Domain-Motif Interaction (DMI) enrichment

Known DMI and SLiM information available in ELM database was used to evaluate enrichment differences in different high-throughput methods. SLiMEnrich **(Chapter 2**) (Idrees, Perez-Bercoff et al. 2018) was used to evaluate enrichment in different PPI datasets.

An estimation of enrichment is done by permutation test which works by randomly selecting proteins to make new random interaction pairs without replacement from the original PPI data. Proteins maintain an identical degree due to the permutation without replacement. Each dataset is permutated 1000x to get better estimation of the random DMIs. Enrichment is calculated as an empirical P-value corresponding to the probability of seeing at least as many DMIs in random PPI data. DMI enrichment (E-score) was calculated as the ratio of the number of predicted DMI to the mean random DMI. Total proportion of potential DMI found in PPIs was calculated, i.e. the total proportion of those DMI that were theoretically identifiable given the proteins in the PPI datasets. The distribution of the real DMI count over 1000x randomisations, $\mathrm{DMI}_{\mathrm{Real}}$, was estimated as follows:

$$\mathrm{DMI}_{\mathrm{Real}} = \mathrm{DMI}_{\mathrm{Obs}} - \mathrm{DMI}_{\mathrm{Ran}}$$

Where $\mathrm{DMI}_{\mathrm{Obs}}$ is the number of observed DMIs in the real PPI dataset and $\mathrm{DMI}_{\mathrm{Ran}}$ is the distribution of observed DMIs in the random PPI datasets.

Normalisation of the data was done by dividing number of real DMIs by mean random DMIs.

For this analysis, ELMi-Protein strategy of SLiMEnrich **(Figure 2.1)** was employed to evaluate enrichment in different publicly available datasets. ELMi-Protein strategy works by mapping PPI protein pairs directly on to known DMI data in ELM. Moreover, impact of each ELM type on enrichment was observed to see if there were any ELM types that were making DMI enrichment better or worse in different high-throughput methods. A Pearson's pairwise chi-square test was done to see how significant different datasets were in comparison to others in terms of capturing different sorts of interactions. For this purpose, pairwise comparisons of each possible combination of datasets were done and p-value was calculated.

3.3.3 DMI prediction quality

Different types of DMI data was used to assess quality of predictions. First, ELMc-Protein strategy was used where enrichment was calculated using known ELM instances available in ELM. Noise in the DMI prediction was increased by adding domain information in the DMI network. For this purpose, ELMc-Domain strategy was used where known ELM instances were mapped to their Pfam domain partners.

3.3.4 Domain-Domain Interaction (DDI) enrichment

Domain-Domain Interaction (DDI) enrichment was also evaluated where we used experimentally validated DDI data from 3DID on 2018-11-28 (Mosca, Ceol et al. 2014) (https://3did.irbbarcelona.org/download.php). 3DID is a databases of high-resolution 3D structures of known PPIs which makes it a highly reliable resource of known interactions (Mosca, Ceol et al. 2014). PDB Ids of 3D DDI complexes and their interacting chain information was extracted from the 3DID database. The interacting PDB chains were then mapped to their corresponding Uniprot proteins. PDB chains to Uniprot mapping was done using PDBSWS (Martin 2005). DDI protein pairs were made non-redundant and were restricted to reviewed Uniprot proteins only. The resulting DDI pairs (5,589 DDI) were then used as known DDI dataset to evaluate enrichment in different datasets.

3.3.5 PPI prediction quality

Enrichment was used to evaluate the impact of PPI quality. HIPPIE PPIs were grouped into different subsets based on their confidence scores (0-1). Eleven subsets were generated where each subset had PPIs from certain confidence scores: Subset 0 which had PPIs with 0 confidence score, 0.1 which had PPIs ranging from 0.11-0.19 confidence scores, 0.2 which had PPIs ranging from 0.21-0.29 confidence scores, 0.3 which had PPIs ranging from 0.31-0.39 confidence scores, 0.4 which had PPIs ranging from 0.41-0.49 confidence scores, 0.5 which had PPIs ranging from 0.51-0.59 confidence scores, 0.6 which had PPIs ranging from 0.61-0.69 confidence scores, 0.7 which had PPIs ranging from 0.71-0.79 confidence scores, 0.8 which had PPIs ranging from 0.81-0.89 confidence scores, 0.9 which had PPIs ranging from 0.91-0.99 confidence scores and subset 1 which had PPIs having 1 as their confidence score.

3.4 Results

3.4.1 Enrichment analysis

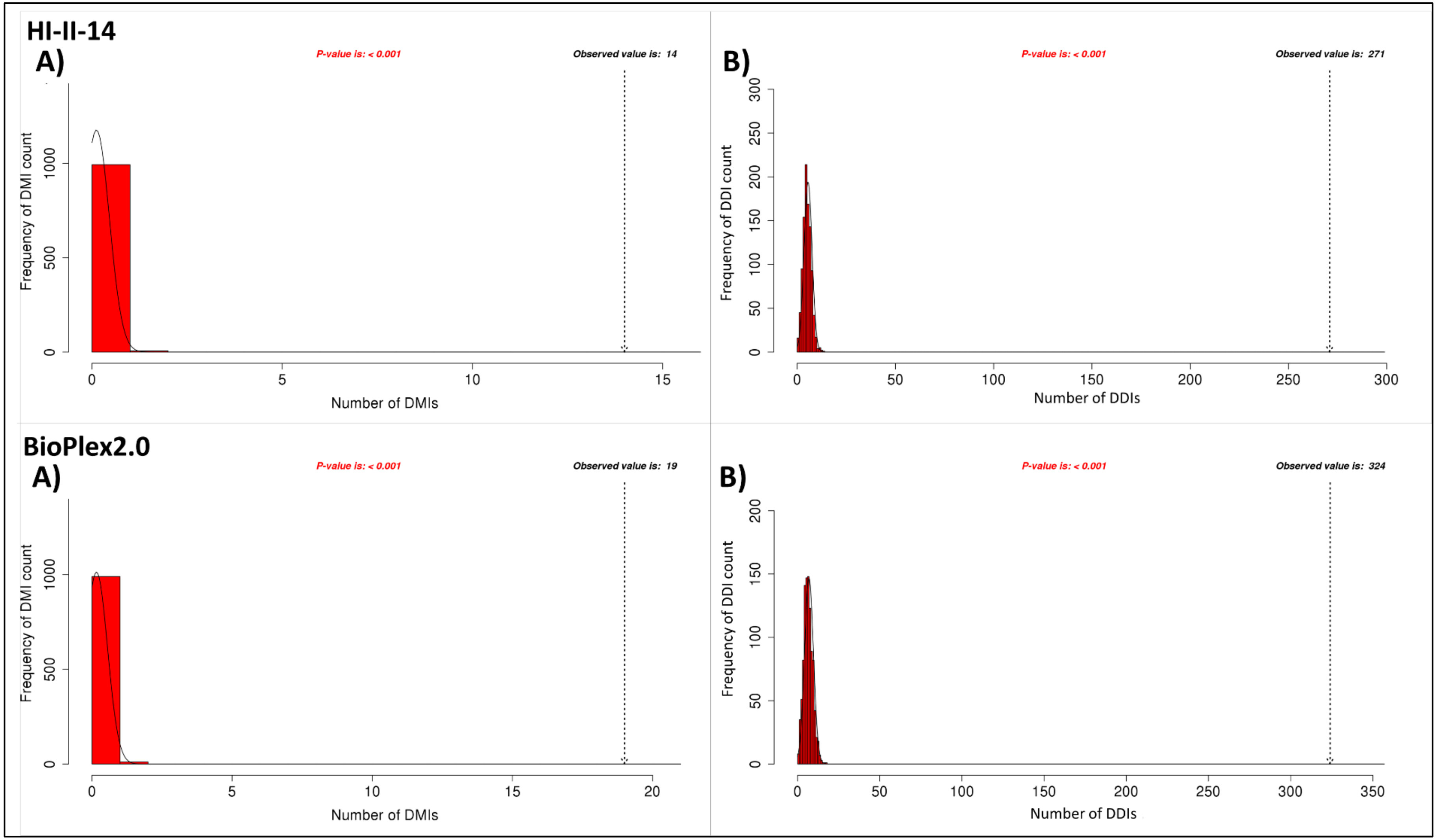
In this study, we have compared five different proteome-wide human interactomes and five publicly available databases to evaluate their efficiency as a source of capturing Domain-Motif Interactions (DMIs) and Domain-Domain Interactions (DDIs). First, we extracted human interactions from BioGrid (409,173) of which 304,409 (74%) were successfully mapped to Uniprot IDs. CoFrac-12 had 13,918 PPIs of which 13,849 (99.5%) were mapped to Uniprot IDs. CoFrac-15 had 16,655 PPIs of which 16,287 (97.7%) were mapped to Uniprot IDs. HI-II-14 had 13,945 PPIs of which 13,410 (96.1%) were mapped to Uniprot IDs. HPRD had 39,240 PPIs of which 37,203 (94.8%) were mapped to Uniprot IDs. HIPPIE had 340,629 PPIs of which 301,235 (88.4%) were mapped to Uniprot IDs. Similarly, BioGrid had 409,173 PPIs of which 304,409 (74.3%) were successfully mapped to Uniprot IDs **(Table 3.1, Figure 3.1).**

Table 3.1. Comparison of different PPI datasets as a source of capturing DMIs and DDIs.

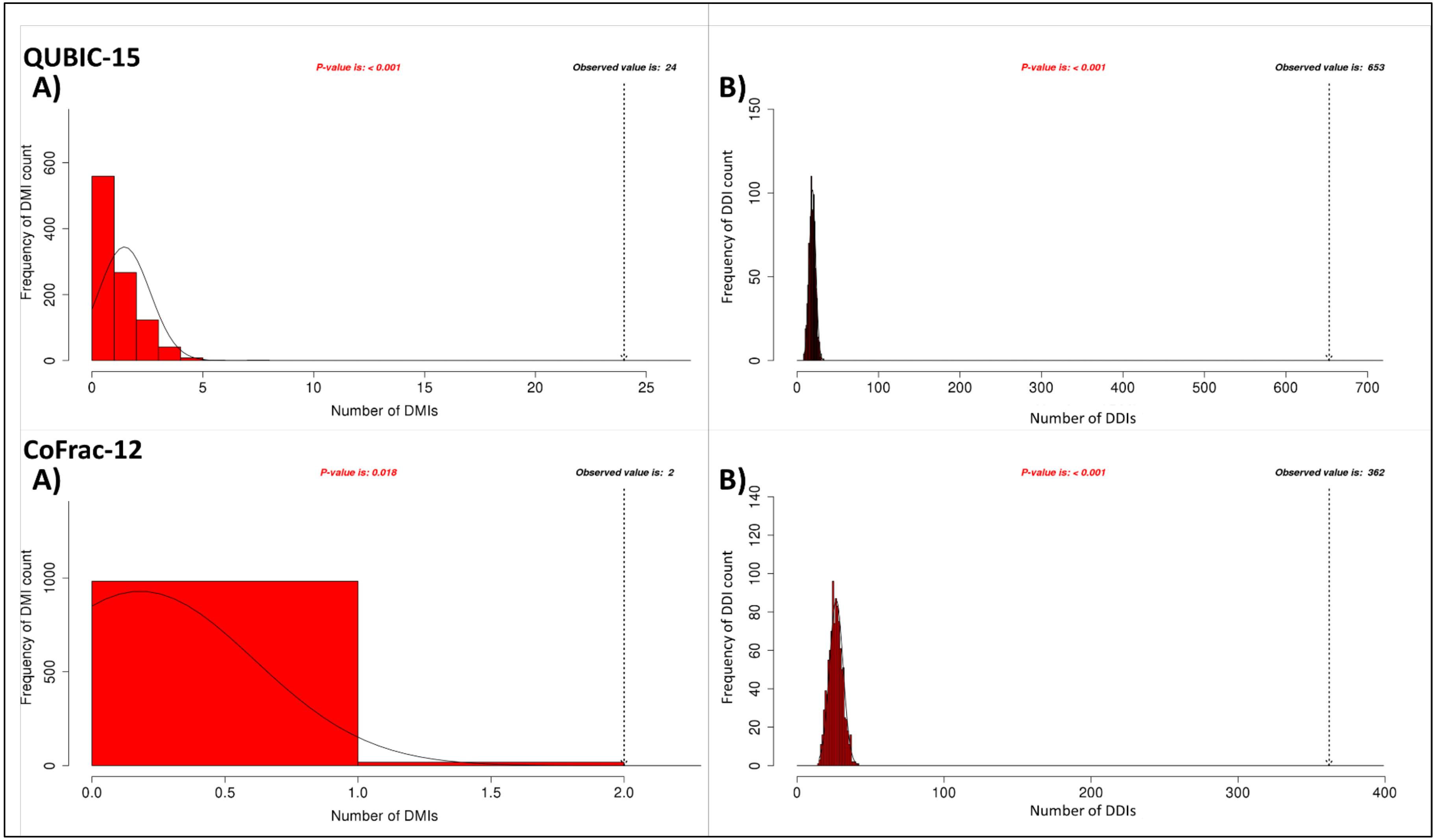
Dataset	**PPIs**[1]	**Method**	**potDMIs**[2]	**DMIs**[3]	**DMI Enrichment**[4] **(3 s.f)**	**potDDIs**[5]	**DDIs**[6]	**DDI Enrichment**[7] **(3 s.f.)**
HI-II-14 **(Rolland, Tasan et al. 2014)** **[retrieved: 2016-06-17]**	25,956	Y2H	61	14	121**	1,272	271	49.0**
BioPlex2.0 **(Huttlin, Bruckner et al. 2017)** **[retrieved: 2017-05-29]**	53,710	AP-MS	137	19	120**	2,296	324	46.8**
QUBIC-15 (Hein, Hubner et al. 2015) **[retrieved: 2017-05-17]**	50,573	AP-MS	186	24	16.4**	2,149	653	34.4**
CoFrac-12 (Havugimana, Hart et al. 2012) **[retrieved: 2018-06-05]**	27,643	CoFrac-MS	63	2	11.0*	1,544	362	13.5**
CoFrac-15 **(Wan, Borgeson et al. 2015)** **[retrieved: 2017-05-31]**	32,452	CoFrac-MS	91	6	23.9**	1,810	395	13.7**
HPRD (Keshava Prasad, Goel et al. 2009) **[retrieved: 2018-11-13]**	71,811	All	493	236	23.2**	4,245	1,900	57.7**
HINT (Das and Yu 2012) **[retrieved: 2018-11-13]**	81,788	All	389	41	21.4**	3,559	536	23.0**
IntAct (Hermjakob, Montecchi-Palazzi et al. 2004) **[retrieved: 2018-11-13]**	159,377	All	526	203	17.2**	4,695	1,332	28.7**
BioGrid (Oughtred, Chatr-Aryamontri et al. 2016) **[retrieved: 2018-11-13]** **v3.5.166**	556,695	All	562	359	23.6**	5,219	2,517	20.9**
HIPPIE (Alanis-Lobato, Andrade-Navarro et al. 2017) **[retrieved: 2018-11-13]** **v2.1**	598,158	All	556	407	17.7**	5,195	3,164	25.0**

*P-value < 0.05, **P-value < 0.001

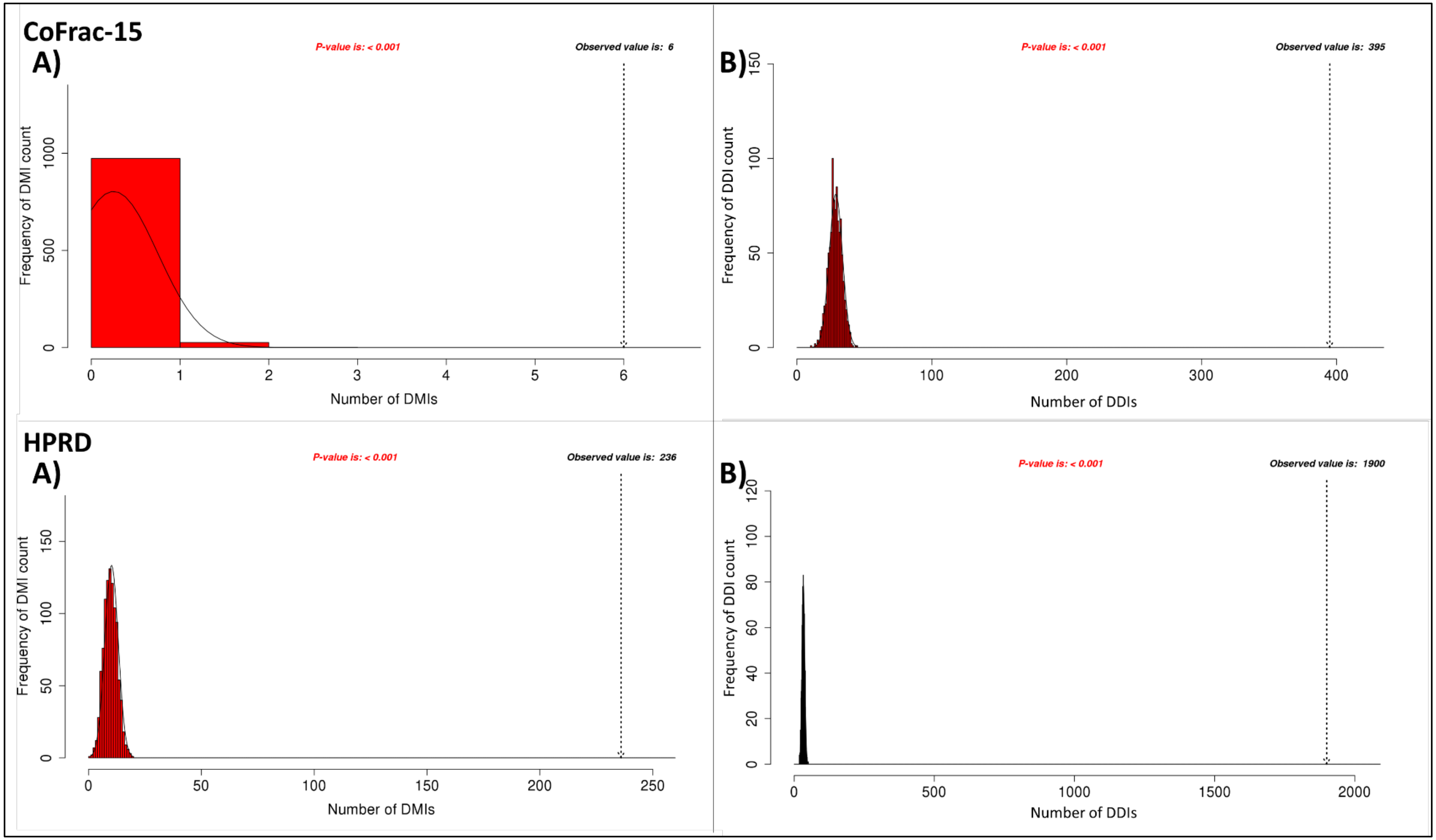
1. Number of symmetrical and non-redundant PPIs having Uniprot reviewed protein pairs.
2. Number of all possible DMIs, given the proteins in each dataset.
3. Known SLiM-Protein interactions from the ELM database (proportion (%) captured from potential DMIs).
4. Observed enrichment of known DMIs captured from PPIs.
5. Number of all possible DDIs, given the proteins in each dataset.
6. Known DDIs from the 3did database (proportion (%) captured from potential DDIs).
7. Observed enrichment of known DDIs captured from PPIs.


HI-II-14
A)
P-value is: < 0.001
Observed value is: 14
Frequency of DMI count
0
500
1000
0
5
10
15
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 271
Frequency of DDI count
0
50
100
150
200
250
300
0
50
100
150
200
250
300
Number of DDIs
BioPlex2.0
A)
P-value is: < 0.001
Observed value is: 19
Frequency of DMI count
0
500
1000
0
5
10
15
20
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 324
Frequency of DDI count
0
50
100
150
200
0
50
100
150
200
250
300
350
Number of DDIs

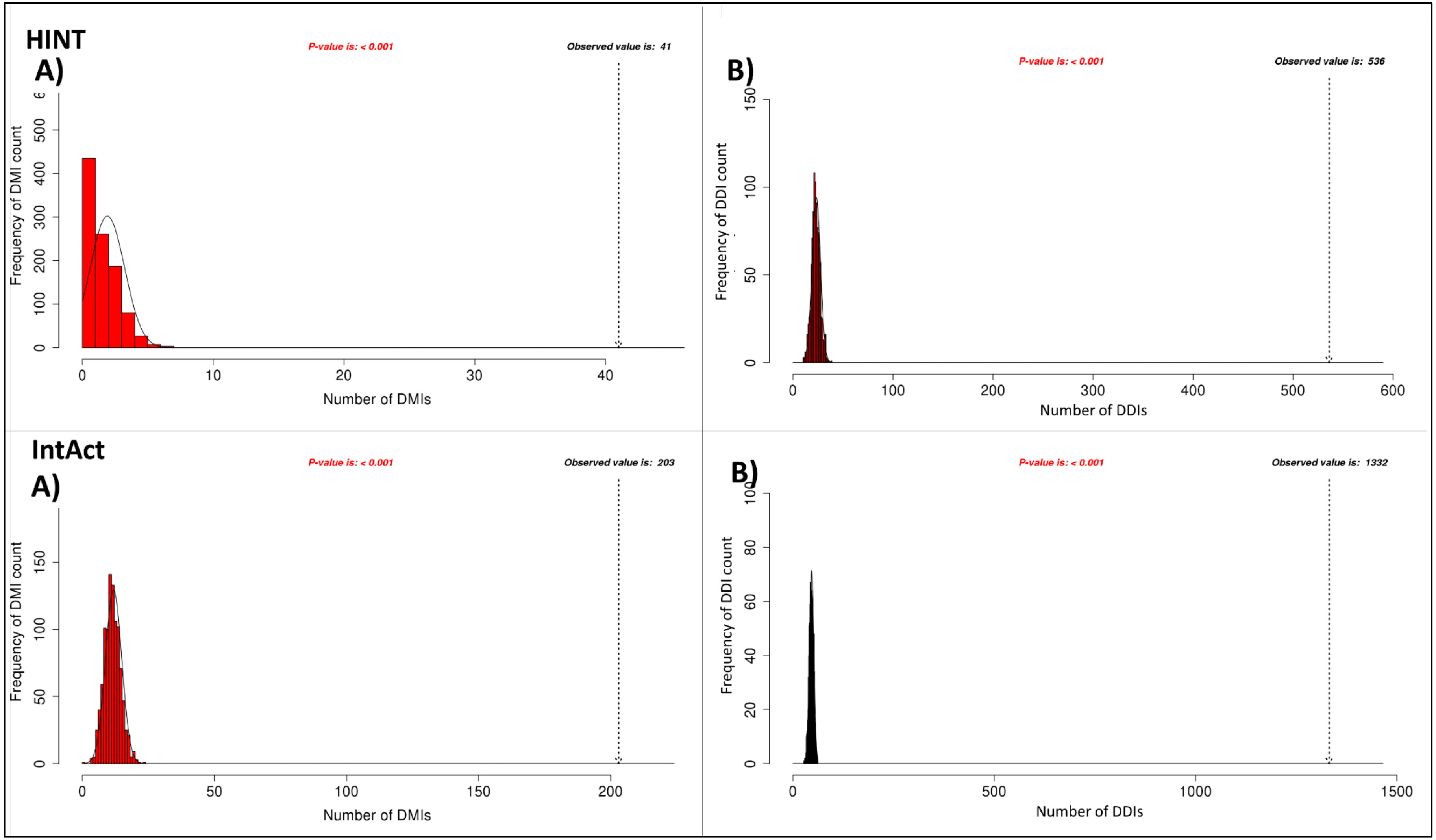
Cont.


QUBIC-15
A)
P-value is: < 0.001
Observed value is: 24
Frequency of DMI count
0
200
400
600
0
5
10
15
20
25
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 653
Frequency of DDI count
0
50
100
150
0
100
200
300
400
500
600
700
Number of DDIs
CoFrac-12
A)
P-value is: 0.018
Observed value is: 2
Frequency of DMI count
0
500
1000
0.0
0.5
1.0
1.5
2.0
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 362
Frequency of DDI count
0
20
40
60
80
100
120
140
0
100
200
300
400
Number of DDIs

Cont.


CoFrac-15
A)
P-value is: < 0.001
Observed value is: 6
Frequency of DMI count
0
500
1000
0
1
2
3
4
5
6
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 395
Frequency of DDI count
0
50
100
150
0
100
200
300
400
Number of DDIs
HPRD
A)
P-value is: < 0.001
Observed value is: 236
Frequency of DMI count
0
50
100
150
0
50
100
150
200
250
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 1900
Frequency of DDI count
0
20
40
60
80
100
120
0
500
1000
1500
2000
Number of DDIs

Cont.


HINT
A)
P-value is: < 0.001
Observed value is: 41
Frequency of DMI count
0
100
200
300
400
500
6
0
10
20
30
40
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 536
Frequency of DDI count
0
50
100
150
0
100
200
300
400
500
600
Number of DDIs
IntAct
A)
P-value is: < 0.001
Observed value is: 203
Frequency of DMI count
0
50
100
150
0
50
100
150
200
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 1332
Frequency of DDI count
0
20
40
60
80
100
0
500
1000
1500
Number of DDIs

Cont.

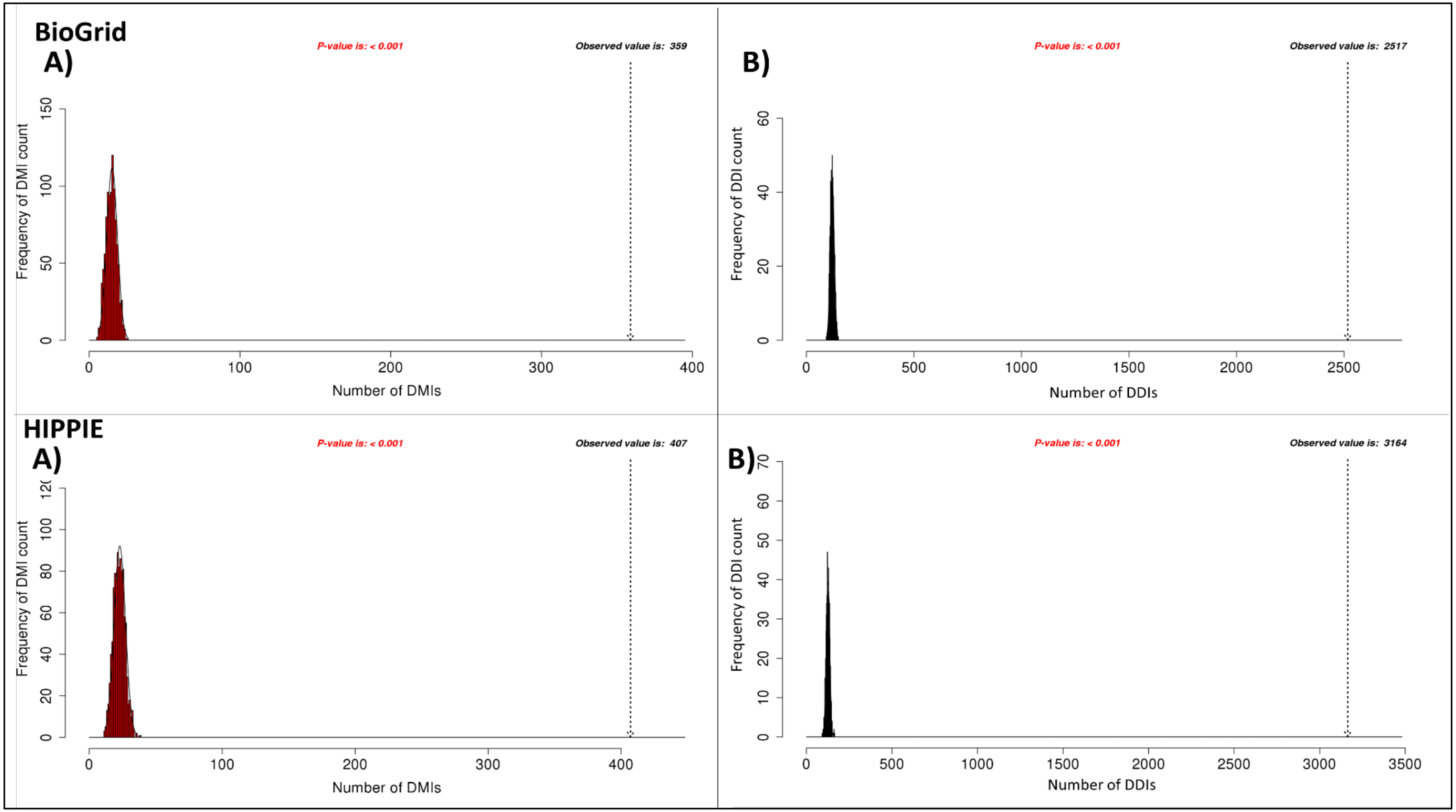


Figure 3.1. Enrichment statistics and histogram of expected random DMI and DDI counts in HI-II-14 dataset using known data,

A) Absolute number of DMI count in different datasets, **B)** Absolute number of DDI count in different datasets. Frequency bars indicate the number of randomised PPI datasets returning a given number of known DMIs/DDIs. The dotted arrow indicates the observed number of known DMIs/DDIs.

3.5 DMI enrichment in different datasets

All datasets showed significant enrichment (P-value < 0.05) suggesting that they all capture domain-motif interactions**.** The BioPlex2.0 AP-MS dataset had 53,710 symmetrical and non-redundant PPIs, of which only 19 were among the known DMIs in ELM. At face value, this appears to be a disappointing performance. However, permutation testing reveals that this is an enrichment of approx. 120x the expected number of known DMIs that would be captured if the 53,710 PPIs were randomly associated (*P* <0.001). On the other hand, the second AP-MS dataset, QUBIC-15, captured 24 known DMIs in 50,573 PPIs, but enrichment was quite low in comparison to BioPlex2.0. HI-II-14 which has PPIs predicted by Y2H method had 14 known DMIs. Enrichment was similar to BioPlex2.0 121x showing that Y2H screen was capturing DMIs. CoFrac-15 captured 6 known DMIs and was more enriched than QUBIC-15 dataset that captured 24 known DMIs.

To further assess the capability of PPI data for capturing DMIs, we extended our analysis to different comprehensive PPI databases which have data from high-throughput studies. Here, we have selected five well known databases i.e. IntAct, BioGrid, HPRD, HINT and HIPPIE to predict DMIs for human interactome. BioGrid was ranked as highest in terms of enrichment followed by HPRD and HINT **(Figure 3.2A).** HIPPIE captured more DMIs than other databases but had lowest enrichment. The enrichment score of all other datasets was lower than HI-II-14 and BioPlex2.0 **(Figure 3.2A).**

A binary vs co-complex PPI analysis was also done where we extracted binary PPIs (i.e. Y2H) and co-complex PPIs (i.e. AP-MS and CoFrac-MS) from BioGrid, IntAct and HIPPIE databases to evaluate enrichment. Both binary and co-complex PPIs showed significant DMI enrichment when compared with random pairs of proteins. Both Y2H and AP-MS captured significant number of known DMIs while CoFrac-MS didn't perform well **(Table 3.3, Figure 3.2A).**

3.6 DDI enrichment in different datasets

We also checked how good these datasets were in terms of capturing domain-domain interactions (DDIs). For this purpose, we used known DDI data from 3did database (Mosca, Ceol et al. 2014) and evaluated enrichment in different datasets **(Figure 3.2B).**

HIPPIE captured highest number of DDIs as compared to other datasets. All datasets captured DDIs with significant enrichment (p-value < 0.001). Among high-throughput screens, HI-II-14 showed highest enrichment followed by BioPlex2.0 and QUBIC-15. Both CoFrac datasets also showed significant DDI enrichment. Among databases, HPRD was the most enriched dataset in terms of capturing DDIs followed by IntAct, HIPPIE, HINT and BioGrid **(Figure 3.2B).**

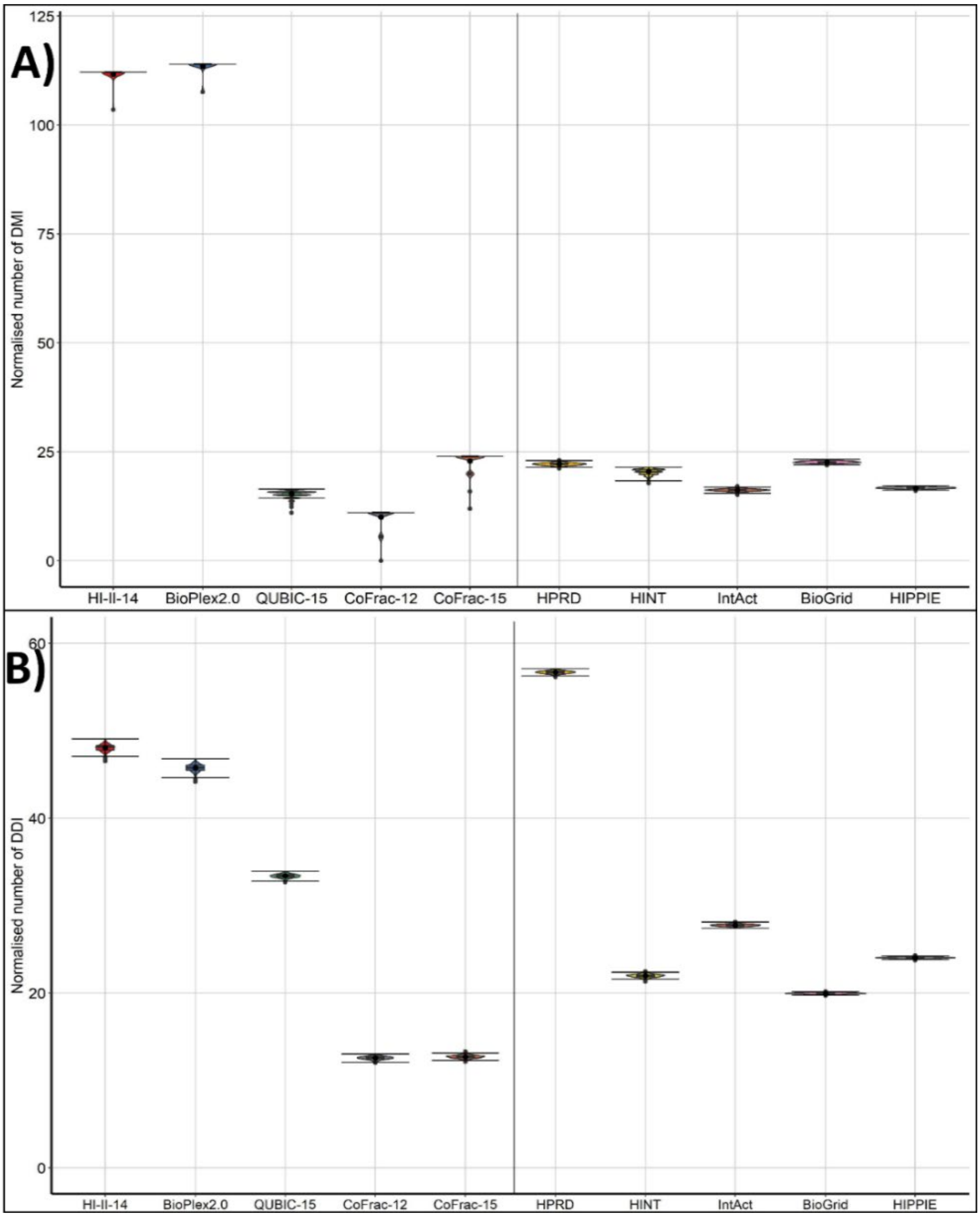


Figure 3.2. Normalised number of DMIs and DDIs captured by different datasets.

A) Normalised number of DMIs captured over 1000x randomisations. Y-axis is the normalized number of DMIs and each bar represents number of real DMIs captured over 1000x randomisations by subtracting random DMIs from observed DMIs. Left panel shows DMIs captured by high-throughput methods and right panel shows DMIs captured by databases, **B)** Normalised number of DDIs captured over 1000x randomisations. Y-axis is the normalized number of DDIs and each bar represents number of real DDIs captured over 1000x randomisations by subtracting random DDIs from observed DDIs. Left panel shows DDIs captured by high-throughput methods and right panel shows DDIs captured by databases.

3.6.1 Proportion of DMI vs DDI being captured

Total proportion of the DMIs captured from known DMIs showed that HIPPIE captured highest proportion (52%) of DMIs from known human DMIs (789 DMIs) followed by BioGrid which captured 46% of DMIs. All other datasets captured lower proportion of DMI as compared to these two **(Figure 3.3A).**

Total proportion of the DDIs captured from total known DDIs showed that HIPPIE captured highest proportion (~56%) of DDIs from total known DDIs (5,589 DDIs) followed by BioGrid which captured 45% of DDIs, HPRD which captured 34% DDIs and IntAct which captured 23% DDIs. All other datasets captured lower proportion of DDI as compared to these datasets **(Figure 3.3B).**

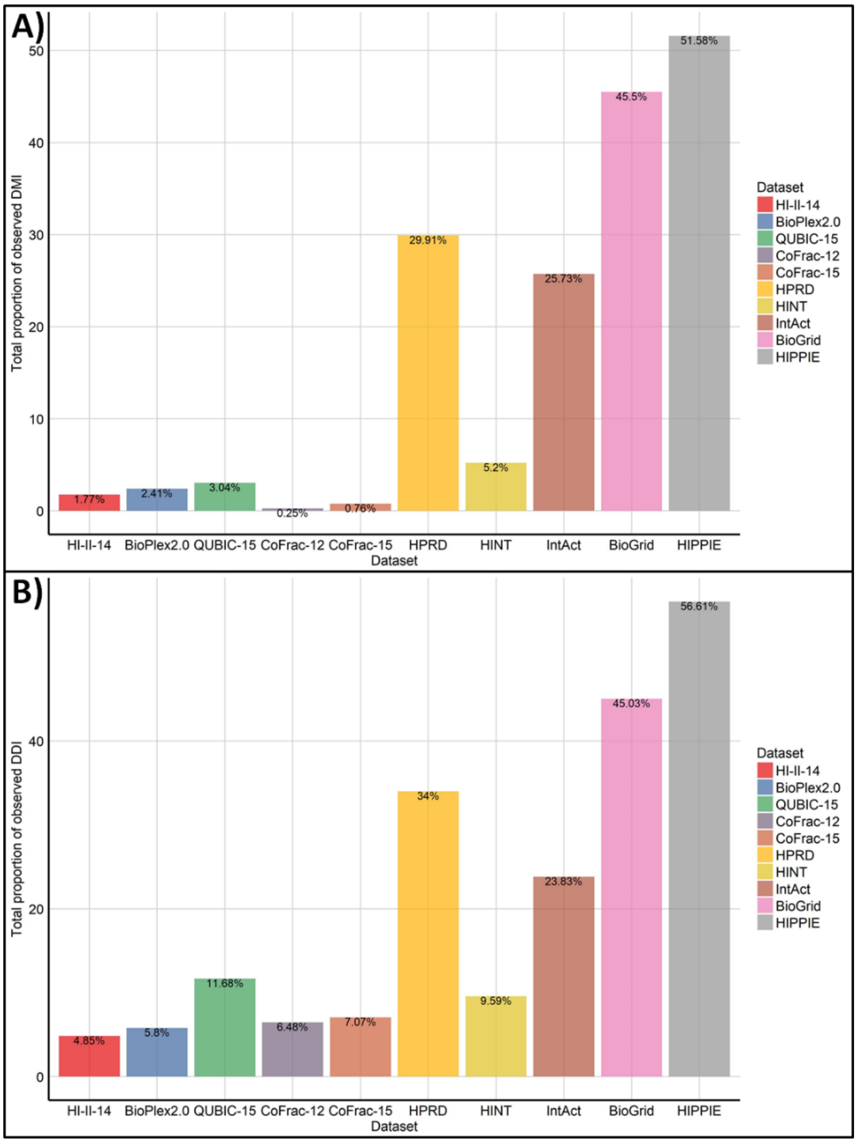


Figure 3.3. Total proportion of DMIs and DDIs captured from known datasets.

A) DMIs captured from known DMI dataset (789 DMIs). Y-axis shows the percentage of DMIs being captured from known DMI dataset. **B)** DDIs captured from known DDI dataset (5,589 DDIs). Y-axis shows the percentage of DDIs being captured from known DDI dataset and ** represents significance at p-value < 0.001

Looking at our analysis, <1% DMIs and < 3% DDIs can be explained by the known DMIs and DDIs. Moreover, there aren't many PPIs that are discovered to be both DMIs and DDIs (< 0.02%) **(Figure 3.4).** A Chi-square pairwise test was done to see significance of each dataset in relation to other datasets **(Table 3.2).**

Table 3.2. Comparison of known DMIs/DDIs captured by different datasets.

Interaction type	**Enriched dataset vs others**
DMI	HPRD > IntAct**
	HPRD > HIPPIE**
	HPRD > HINT**
	HPRD > BioGrid**
	HPRD > QUBIC-15**
	HPRD > CoFrac-15**
	HPRD > CoFrac-12**
	HPRD > BioPlex2.0**
	HPRD > HI-II-14**
	IntAct > HIPPIE**
	IntAct > HINT**
	IntAct > BioGrid**
	IntAct > BioPlex2.0**
	IntAct > CoFrac-12**
	HIPPIE > CoFrac-12**
	HINT > CoFrac-12*
	HI-II-14 > CoFrac-12*
	BioGrid > CoFrac-12 **
	IntAct > CoFrac-15**
	HINT > CoFrac-15*
	BioGrid > CoFrac-15*
	IntAct > HI-II-14*
	BioGrid > BioPlex2.0*
DDI	**Enriched dataset vs others**
	HPRD > IntAct**
	HPRD > HIPPIE**
	HPRD > HINT**
	HPRD > BioGrid**
	HPRD > QUBIC-15**
	HPRD > CoFrac-15**
	HPRD > CoFrac-12**
	HPRD > BioPlex2.0**
	HPRD > HI-II-14**
	IntAct > HIPPIE**
	IntAct > HINT**
	IntAct > BioGrid**
	IntAct > BioPlex2.0**
	BioPlex2.0 > HIPPIE*
	CoFrac-12 > IntAct**
	CoFrac-12 > HIPPIE**
	CoFrac-12 > HINT**
	CoFrac-12 > QUBIC-15*
	CoFrac-12 > BioGrid**
	CoFrac-15 > HIPPIE*
	CoFrac-15 > IntAct**
	CoFrac-15 > HINT*
	CoFrac-15 > BioGrid*
	QUBIC-15 > IntAct**
	HI-II-14 > IntAct*

*shows p-value <0.05 and ** shows p-value <0.001.

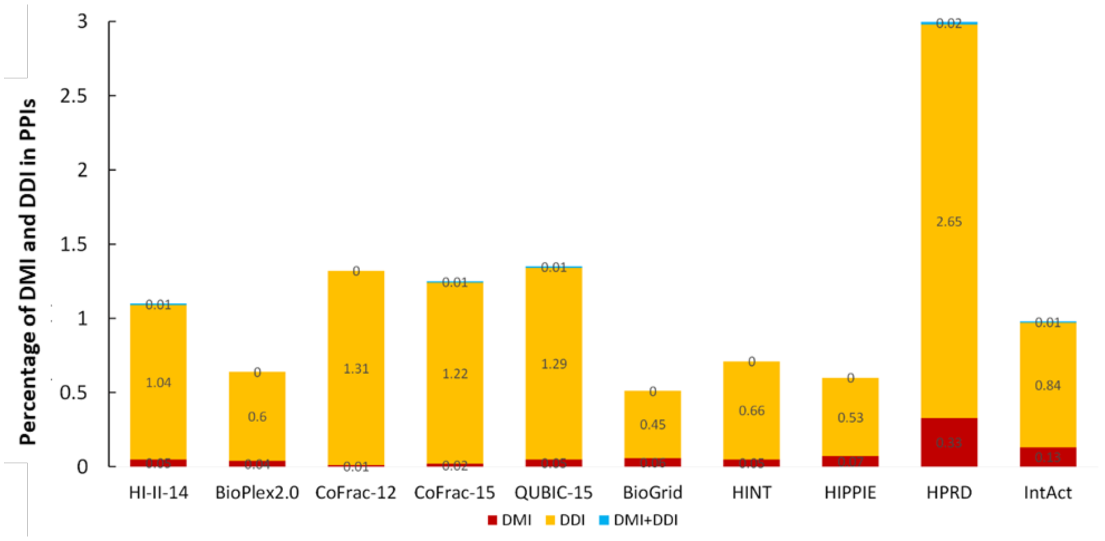


Figure 3.4. Percentage of PPIs that are known DMIs, DDIs or DMIs+DDIs (non-redundant and reviewed proteins only).

Y-axis shows the percentage of PPIs that can be explained as DMIs or DDIs, numbers inside bars shows proportion in percentage from total numbers of PPIs.

3.6.2 High-throughput screens capture DMIs and DDIs

As both BioPlex2.0 and HI-II-14 datasets showed quite high enrichment (~120x) suggesting that Y2H and AP-MS screens were indeed capturing DMIs, we decided to further investigate which method among Y2H and AP-MS was better at capturing DMIs. For this purpose, we pulled out PPIs identified by two-hybrid, AP-MS and CoFrac-MS from three well known PPI databases i.e. BioGrid, IntAct and HIPPIE and used them to evaluate enrichment **(Figure 3.5A).**

We also checked which method among AP-MS, two-hybrid and CoFrac-MS was better at capturing DDIs. Two hybrid detected interactions available in IntAct and HIPPIE showed more enrichment than AP-MS while AP-MS interactions of BioGrid showed better enrichment than two-hybrid interactions. It can be said that both these techniques are effectively capturing DDIs. CoFrac-MS captured DDIs showed lower enrichment than other high-throughput methods **(Table 3.3, Figure 3.5B).**

Table 3.3. Comparison of different high-throughput methods as a source for capturing DMIs and DDIs.

Dataset	Method	PPIs[1]	potDMIs[2]	DMIs[3]	DMI Enrichment[4] (3 s.f.)	potDDIs[5]	DDIs[6]	DDI Enrichment[7] (3 s.f.)
BioGrid	AP-MS	153,530	413	71	42.2**	2,705	688	30.6**
	Two-hybrid	66,882	370	38	26.1**	3,321	455	25.9**
	CoFrac-MS	36,056	170	7	23.0**	2,026	396	18.2**
IntAct	AP-MS	17,938	157	22	18.5**	1,767	184	8.85**
	Two-hybrid	17,541	259	31	33.5**	1,980	220	21.9**
HIPPIE	AP-MS	347,047	534	264	25.8**	4,838	1,535	17.8**
	Two-hybrid	98,283	500	140	46.4**	4,065	925	40.8**
	CoFrac-MS	542	18	1	13.2**	176	30	17.4**

**P-value < 0.001

1. Number of symmetrical and non-redundant PPIs having Uniprot reviewed protein pairs.
2. Number of all possible DMIs, given the proteins in each dataset.
3. Known SLiM-Protein interactions from the ELM database.
4. Observed enrichment of known DMIs captured from PPIs
5. Number of all possible DDIs, given the proteins in each dataset.
6. Known DDIs from the 3did database
7. Observed enrichment of known DDIs captured from PPIs

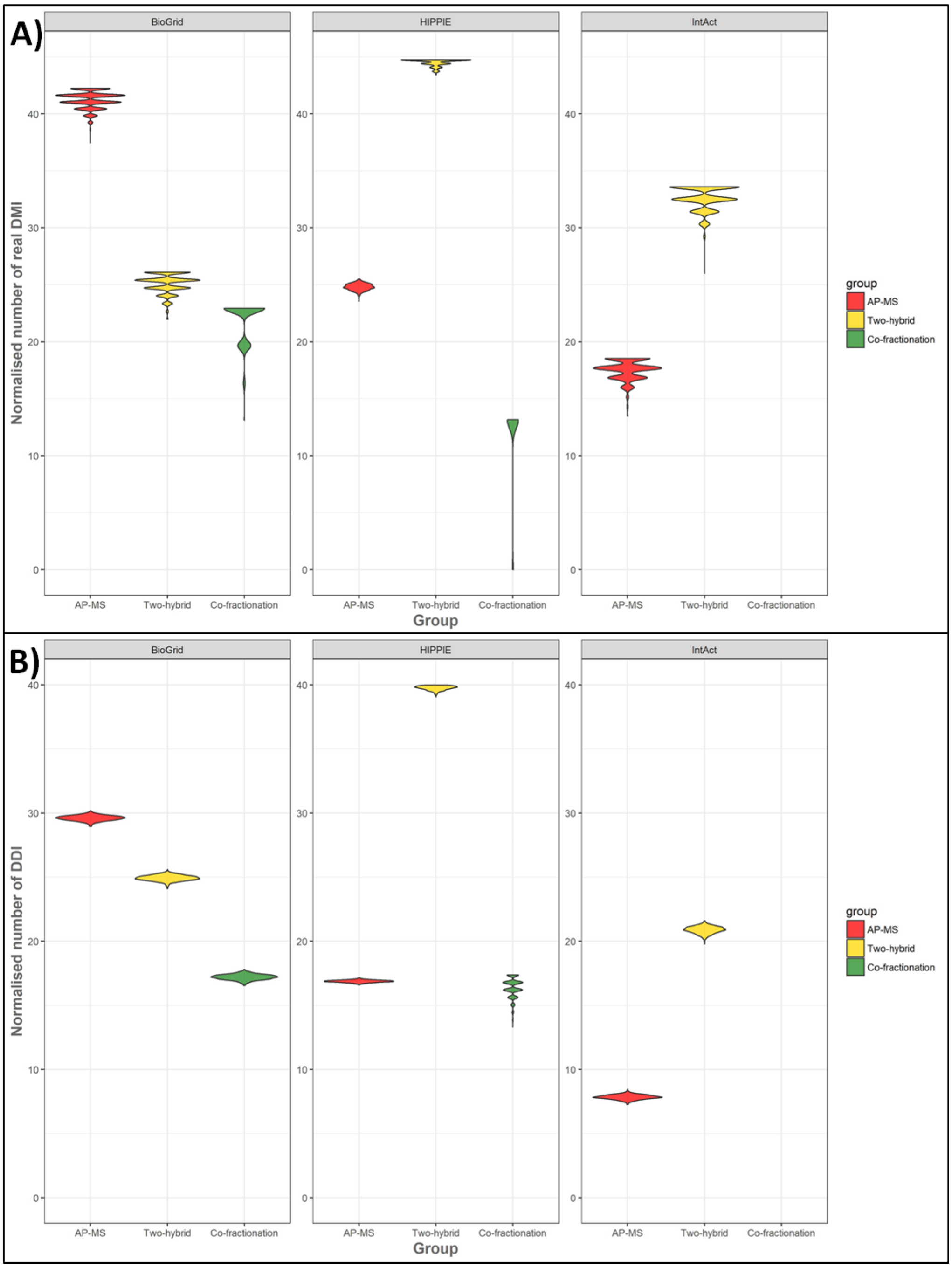


Figure 3.5. Normalised number of real DMIs and DDIs captured by three well known high-throughput methods.

A) Normalised number of real DMIs captured over 1000x randomisations. Red represents DMIs captured by AP-MS, yellow represents DMIs captured by two-hybrid and green represents DMIs captured by co-fractionation screens,

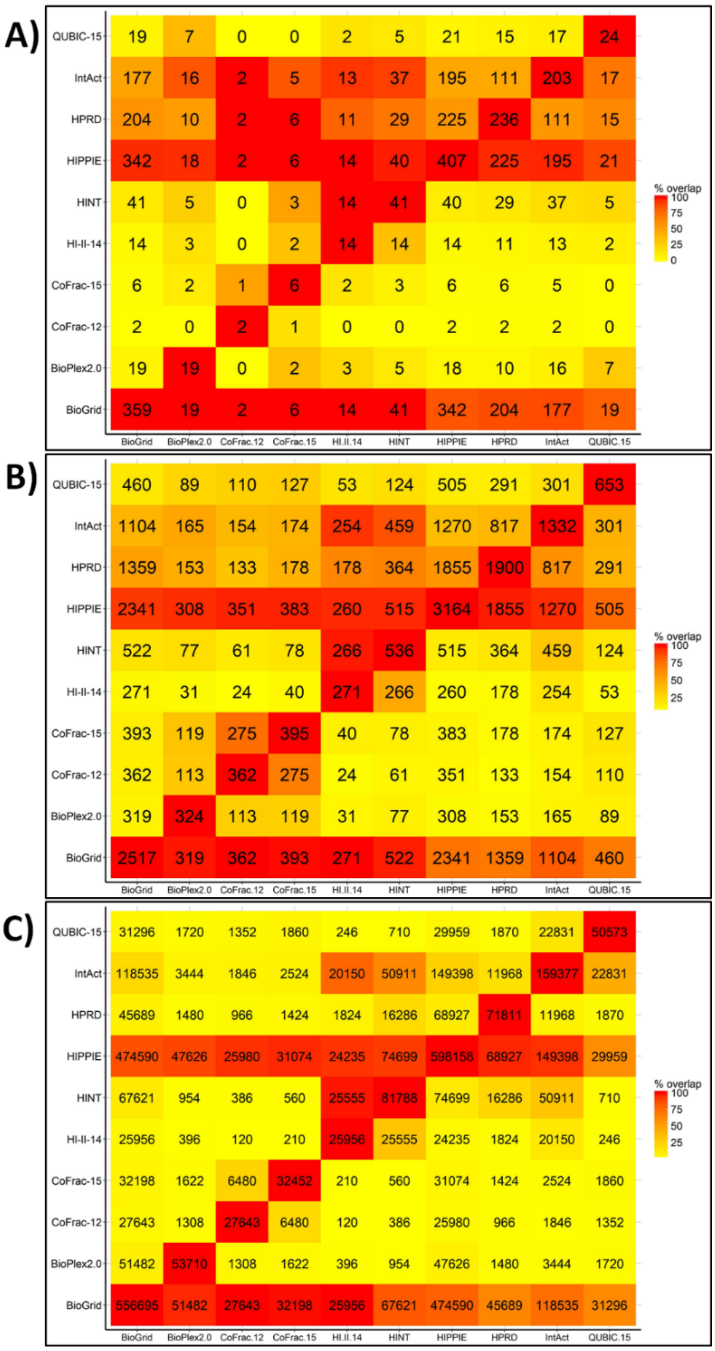
B) Normalised number of real DDIs captured over 1000x randomisations. Red represents DMIs captured by AP-MS, yellow represents DMIs captured by two-hybrid and green represents DMIs captured by co-fractionation screens.

Looking at the enrichment, AP-MS, two-hybrid and CoFrac-MS were capturing DMIs with significant enrichment (P-value < 0.001)**.** Two-hybrid PPIs in HIPPIE database showed highest enrichment than other methods/databases. PPIs identified by AP-MS in BioGrid database also showed better enrichment. Overall, two-hybrid screens were more enriched than other methods in HIPPIE and IntAct databases. Co-fractionation interactions from BioGrid and HIPPIE databases showed lower enrichment than other techniques **(Figure 3.5).**

3.6.3 All PPI resources share overlap of data

Here, we checked overlap of DMIs captured by different high-throughput studies and databases with each other. As expected, most of the DMIs captured by HINT database were overlapping with other databases. All of the DMIs captured by HINT were in the BioGrid database resulting into 100% overlap, 95% overlap with the IntAct database, 97% with HIPPIE and 68% with the HPRD database. Similarly, HIPPIE which is another high-throughput resource showed 85% DMI overlap with BioGrid and 51% overlap with IntAct database. Around 55% of DMIs captured by HIPPIE were in HPRD and 10% of DMI data overlapped with HINT. HPRD also showed significant overlap with other databases i.e. 87% with BioGrid, 49% with IntAct, 13% with HINT and 95% with HIPPIE database **(Figure 3.6A)**. Overall, it can be seen that most of the DMIs captured by HINT, HIPPIE and HPRD are available in BioGrid and IntAct databases.

Looking at the overlaps with IntAct database, HI-II-14 showed 100% overlap, QUBIC had 70%, BioPlex2.0 had 89%, CoFrac-15 had 83% while CoFrac-12 had no overlapping DMIs. On the other hand, all individual high-throughput datasets showed 100% overlap with the BioGrid database except QUBIC which had 79% overlap. Similarly, significant overlap was seen with HIPPIE database where 71% of DMIs captured by HI-II-14 were in the HIPPIE database, 87% of DMI captured by QUBIC, 94% of DMIs captured by BioPlex2.0 while all of the DMIs captured by CoFrac-15 and CoFrac-12 datasets were in HIPPIE. Looking at the comparison with HINT database, all of the DMIs captured by HI-II-14 were in the HINT database while BioPlex2.0 and QUBIC shared small fraction of the DMIs with HINT (26%, 20% respectively). CoFrac-15 had 50% of its DMIs in HINT while CoFrac-12 had none of its two interactions in HINT. Finally, comparison with HPRD showed that HI-II-14 had 78% DMI overlap, QUBIC had 62%, BioPlex2.0 had 52% while CoFrac-15 and CoFrac-12 had 100% overlap with HPRD **(Figure 3.6A).** As expected, most of the DMIs captured by different high-throughput methods and databases were overlapping with each other.



A)

	BioGrid	BioPlex2.0	CoFrac.12	CoFrac.15	HI.II.14	HINT	HIPPIE	HPRD	IntAct	QUBIC.15
QUBIC-15	19	7	0	0	2	5	21	15	17	24
IntAct	177	16	2	5	13	37	195	111	203	17
HPRD	204	10	2	6	11	29	225	236	111	15
HIPPIE	342	18	2	6	14	40	407	225	195	21
HINT	41	5	0	3	14	41	40	29	37	5
HI-II-14	14	3	0	2	14	14	14	11	13	2
CoFrac-15	6	2	1	6	2	3	6	6	5	0
CoFrac-12	2	0	2	1	0	0	2	2	2	0
BioPlex2.0	19	19	0	2	3	5	18	10	16	7
BioGrid	359	19	2	6	14	41	342	204	177	19

B)

	BioGrid	BioPlex2.0	CoFrac.12	CoFrac.15	HI.II.14	HINT	HIPPIE	HPRD	IntAct	QUBIC.15
QUBIC-15	460	89	110	127	53	124	505	291	301	653
IntAct	1104	165	154	174	254	459	1270	817	1332	301
HPRD	1359	153	133	178	178	364	1855	1900	817	291
HIPPIE	2341	308	351	383	260	515	3164	1855	1270	505
HINT	522	77	61	78	266	536	515	364	459	124
HI-II-14	271	31	24	40	271	266	260	178	254	53
CoFrac-15	393	119	275	395	40	78	383	178	174	127
CoFrac-12	362	113	362	275	24	61	351	133	154	110
BioPlex2.0	319	324	113	119	31	77	308	153	165	89
BioGrid	2517	319	362	393	271	522	2341	1359	1104	460

C)

	BioGrid	BioPlex2.0	CoFrac.12	CoFrac.15	HI.II.14	HINT	HIPPIE	HPRD	IntAct	QUBIC.15
QUBIC-15	31296	1720	1352	1860	246	710	29959	1870	22831	50573
IntAct	118535	3444	1846	2524	20150	50911	149398	11968	159377	22831
HPRD	45689	1480	966	1424	1824	16286	68927	71811	11968	1870
HIPPIE	474590	47626	25980	31074	24235	74699	598158	68927	149398	29959
HINT	67621	954	386	560	25555	81788	74699	16286	50911	710
HI-II-14	25956	396	120	210	25956	25555	24235	1824	20150	246
CoFrac-15	32198	1622	6480	32452	210	560	31074	1424	2524	1860
CoFrac-12	27643	1308	27643	6480	120	386	25980	966	1846	1352
BioPlex2.0	51482	53710	1308	1622	396	954	47626	1480	3444	1720
BioGrid	556695	51482	27643	32198	25956	67621	474590	45689	118535	31296

Figure 3.6. Number of DMIs, DDIs and PPIs being overlapped in different publicly available databases.

A) Number of shared DMIs among different datasets, each row and column is a dataset, each box shows the number of shared DMIs, percentage of overlap is shown in gradient where orange represents 100% overlap and yellow represents 0% overlaps, **B)** Number of shared DDIs among different datasets, each row and column is a dataset, each box shows the number of shared DDIs, percentage of overlap is shown in gradient where orange represents 100% overlap and yellow represents 0% overlaps, **C)** Number of shared PPIs among different datasets, each row and column is a dataset, each box shows the number of shared PPIs, percentage of overlap is shown in gradient where orange represents 100% overlap and yellow represents 0% overlaps.

Similarly, we checked what proportion of DDIs was being shared among different datasets. Most of the HINT captured DDIs were also captured by BioGrid (97%), HIPPIE (96%), IntAct (85%) and HPRD (67%) databases. HIPPIE showed 74% DDI overlap with BioGrid, 40% overlap with IntAct database, 59% with HPRD and only 16% with HINT. HPRD showed 97% overlap with HIPPIE, 71% with BioGrid, 43% with IntAct and 19% with HINT database. Just like DMIs, most of the DDIS captured by different high-throughput methods were overlapping with publicly available databases **(Figure 3.6B)**. We also checked proportion of PPIs that was overlapping in different datasets. All high-throughput screens and databases had significant PPI overlap with HIPPIE and BioGrid databases as compared to HINT and HPRD databases **(Figure 3.6C).**

3.6.4 Enrichment drops with the introduction of noise in DMI predictions

As the number of known DMIs captured by high-throughput data was quite low, noisier DMI predictions were used to increase the number of real DMIs identified. The idea was to see if general pattern of enrichment remained same. Here, we first used ELMc-Protein strategy **(Figure 2.1)** where known SLiMs were mapped onto their known protein partners via ELM classes. All datasets remained significantly enriched (P-value < 0.001) over random expectation but the overall enrichment score dropped for all datasets as compared to ELMi-Protein strategy. Enrichment fold of BioPlex2.0 dropped from 120x to 92x and it ranked as the most enriched dataset in comparison to other datasets. On the other hand, enrichment fold of HI-II-14 which was previously 120x dropped to 32x making it second most enriched dataset **(Figure 3.7A).** All other datasets showed lower enrichment than these two datasets. We then used ELMc-Domain strategy where we used domain information. Enrichment for all datasets remained strongly significant (P-value <0.001) **.** Th overall enrichment trend of this strategy remained same as of ELMc-Protein despite a further drop in enrichment scores **(Figure 3.7B).** We also calculated the total proportion of predicted DMIs captured from potential DMIs where potential DMIs shows the total proportion of those DMI that were theoretically identifiable given the proteins in the PPI datasets. HIPPIE and BioGrid databases predicted highest proportion of DMIs from potential DMIs followed by HPRD and IntAct using ELMc-Protein strategy. HINT on the other hand identified only a small proportion of DMIs from potential DMIs. Among high-throughput screens, HI-II-14 identified highest proportion of DMIs followed by BioPlex2.0, QUBIC-15, CoFrac-15 and CoFrac-12 **(Figure 3.9A).** On the other hand, only a small fraction (1-4%) of DMIs was predicted from potential DMIs through ELMc-Domain strategy. The overall trend was same as ELMc-Protein strategy **(Figure 3.9B).**

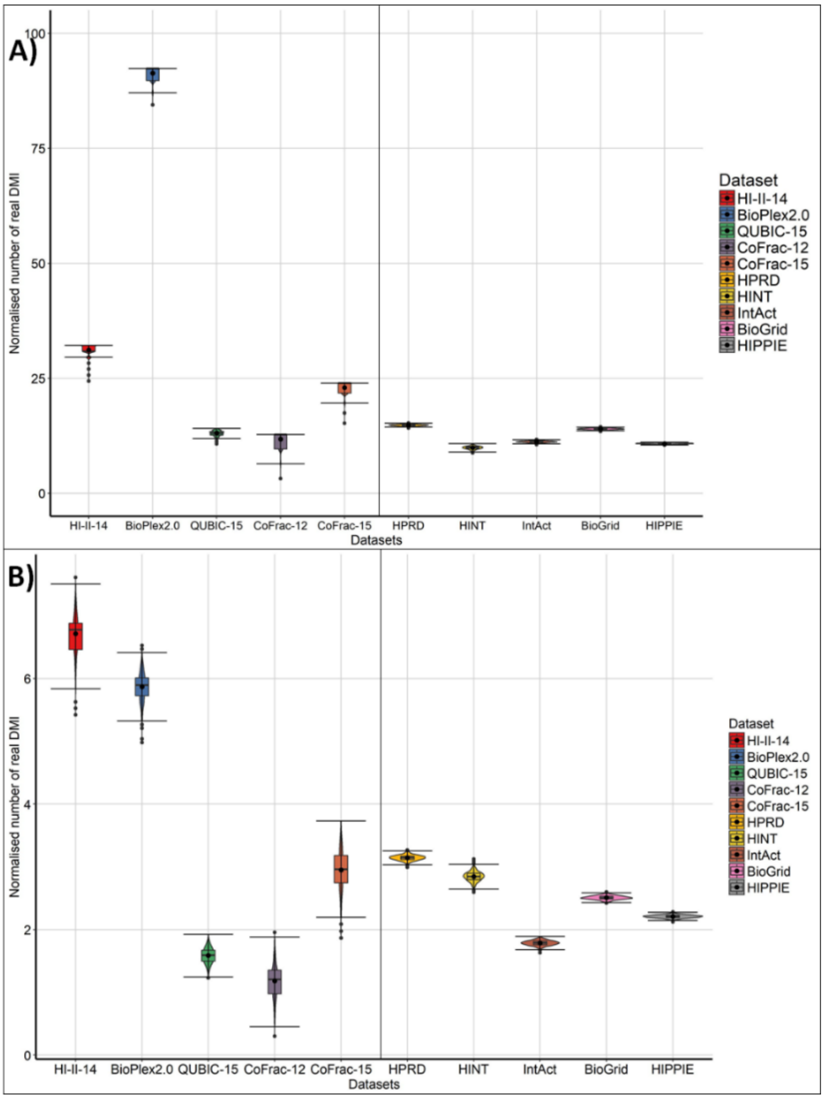
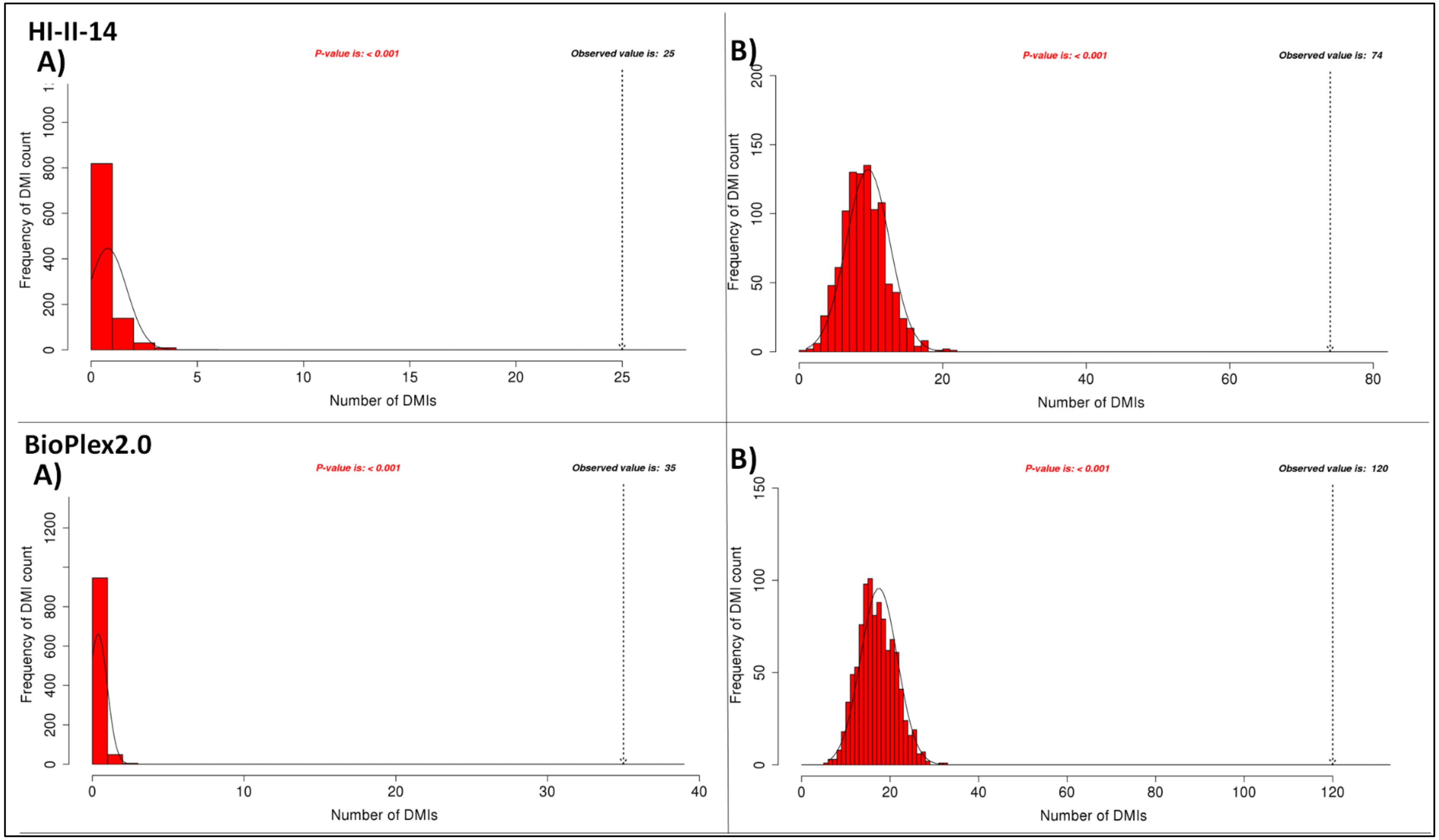


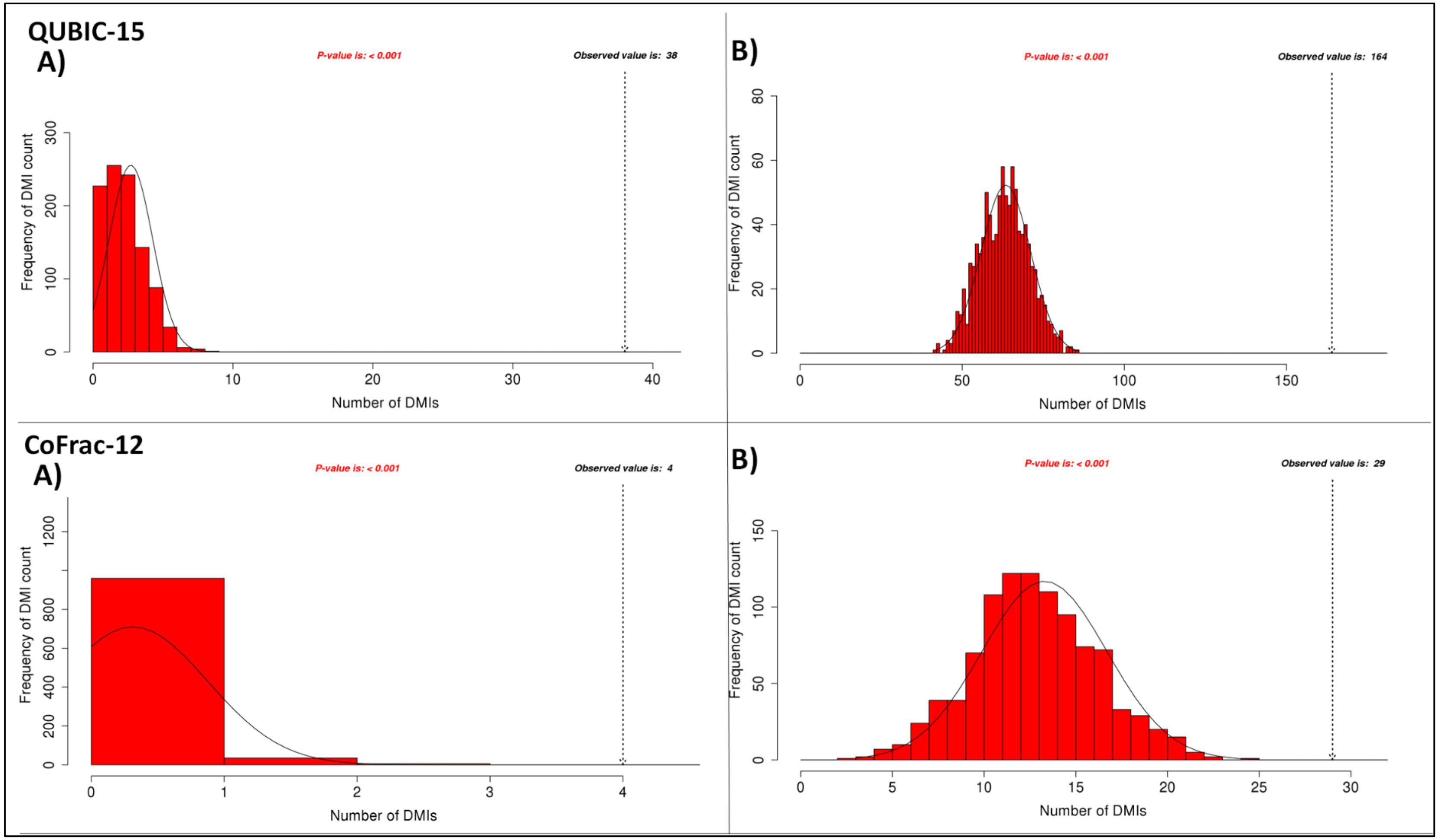
Figure 3.7. Normalised number of DMIs and DDIs captured by different datasets.

A) Normalised number of DMIs captured over 1000x randomisations using ELMc-Protein strategy. Y-axis is the normalized number of DMIs and each bar represents number of real DMIs captured over 1000x randomisations by subtracting random DMIs from observed DMIs. Left panel shows DMIs captured by high-throughput methods and right panel shows DMIs captured by databases, **B)** Normalised number of DMIs captured over 1000x randomisations using ELMc-Domain strategy. Y-axis is the normalized number of DMIs and each bar represents number of real DMIs captured over 1000x randomisations by subtracting random DDIs from observed DMIs. Left panel shows DMIs captured by high-throughput methods and right panel shows DMIs captured by databases.

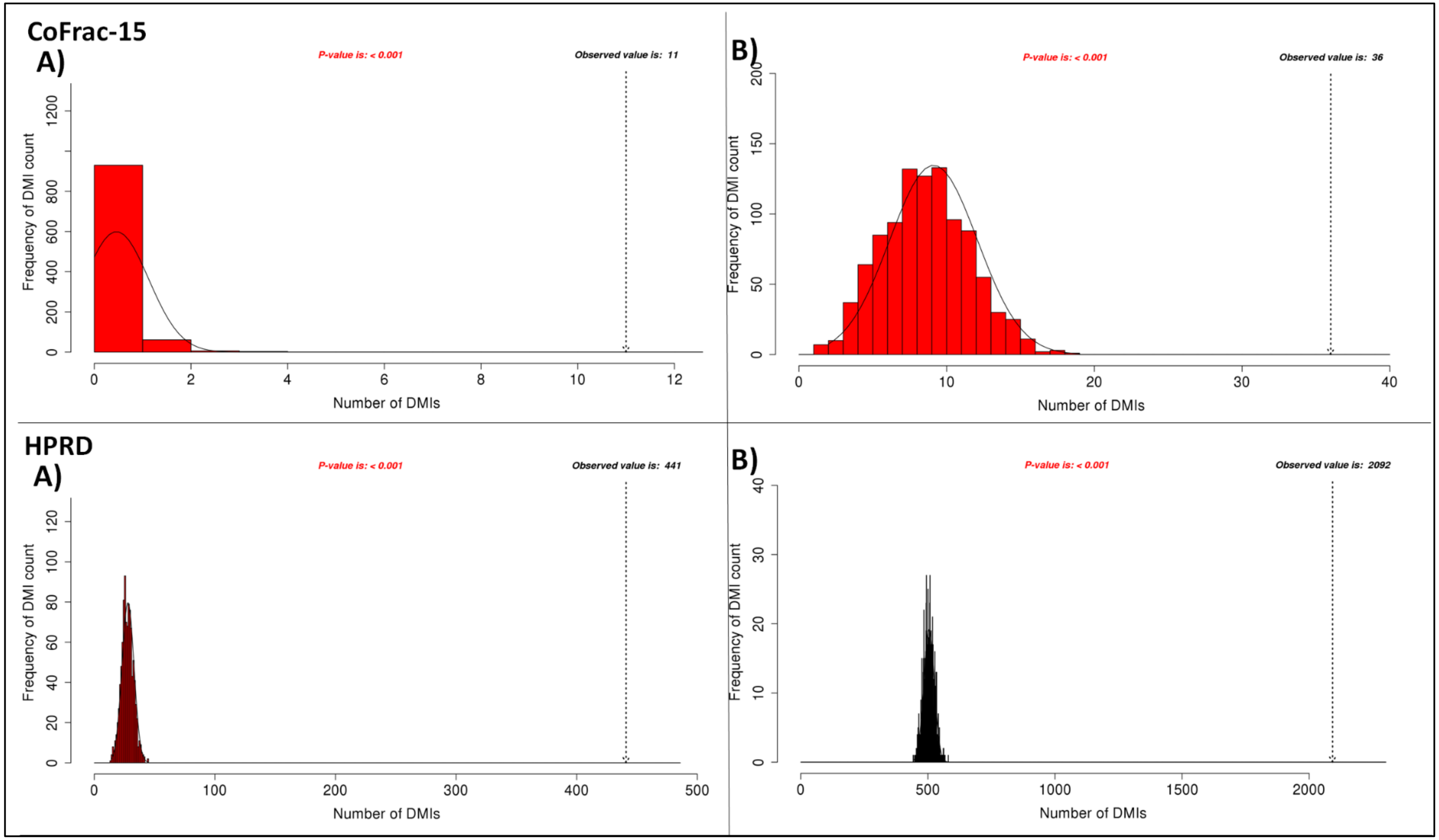
The overall number of DMIs increased with the increase in noise in DMI prediction quality but as the enrichment of these datasets was quite high, the additional DMIs are likely to be real **(Figure 3.8, Figure 3.9)**.


HI-II-14
A)
P-value is: < 0.001
Observed value is: 25
Frequency of DMI count
0
200
400
600
800
1000
1.
0
5
10
15
20
25
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 74
Frequency of DMI count
0
50
100
150
200
0
20
40
60
80
Number of DMIs
BioPlex2.0
A)
P-value is: < 0.001
Observed value is: 35
Frequency of DMI count
0
200
400
600
800
1200
0
10
20
30
40
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 120
Frequency of DMI count
0
50
100
150
0
20
40
60
80
100
120
Number of DMIs

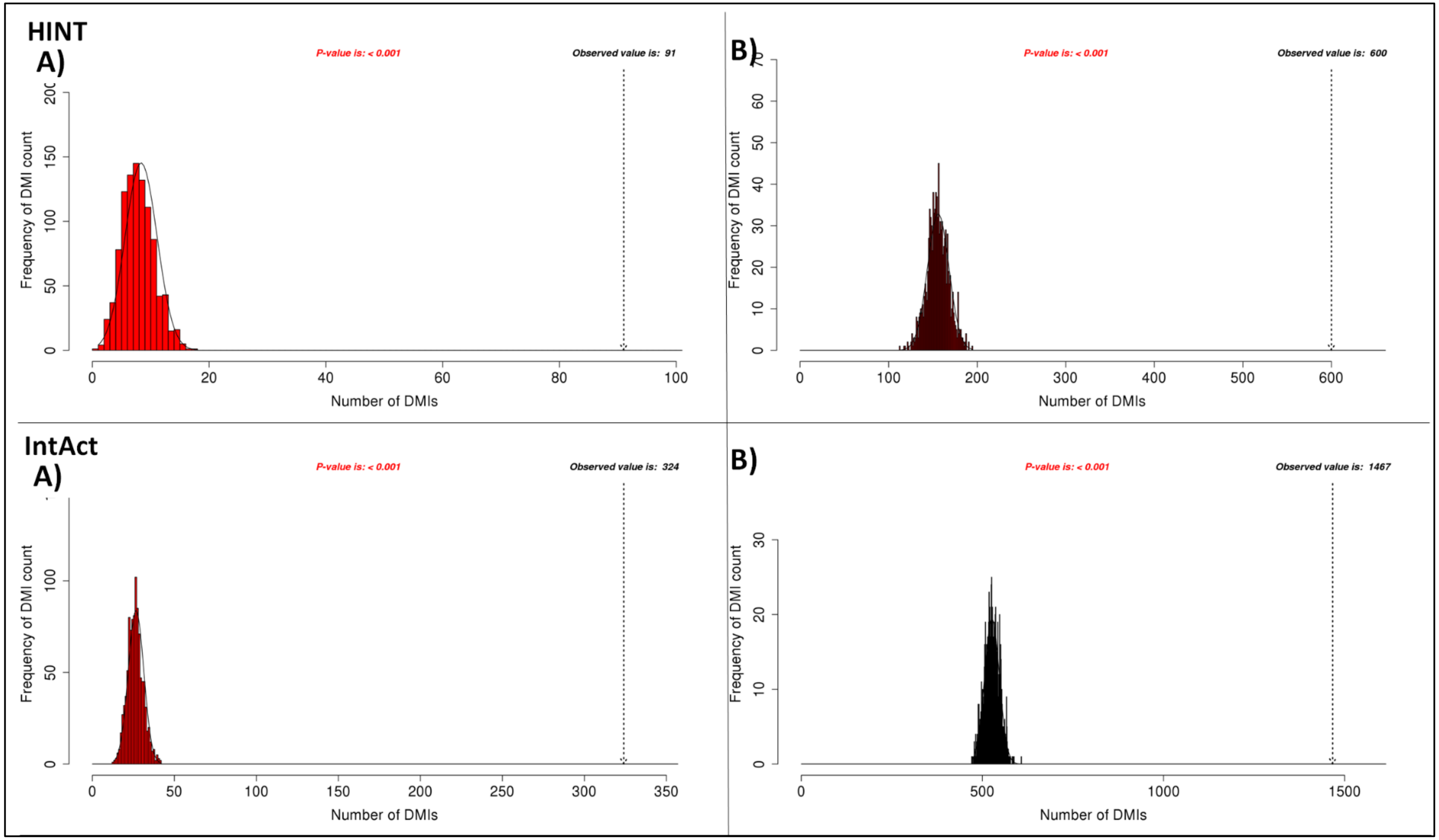
Cont.


QUBIC-15
A)
P-value is: < 0.001
Observed value is: 38
Frequency of DMI count
0
100
200
300
0
10
20
30
40
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 164
Frequency of DMI count
0
20
40
60
80
0
50
100
150
Number of DMIs
CoFrac-12
A)
P-value is: < 0.001
Observed value is: 4
Frequency of DMI count
0
200
400
600
800
1200
0
1
2
3
4
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 29
Frequency of DMI count
0
50
100
150
0
5
10
15
20
25
30
Number of DMIs

Cont.


CoFrac-15
A)
P-value is: < 0.001
Observed value is: 11
Frequency of DMI count
0
200
400
600
800
1200
0
2
4
6
8
10
12
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 36
Frequency of DMI count
0
50
100
150
200
0
10
20
30
40
Number of DMIs
HPRD
A)
P-value is: < 0.001
Observed value is: 441
Frequency of DMI count
0
20
40
60
80
100
120
0
100
200
300
400
500
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 2092
Frequency of DMI count
0
10
20
30
40
0
500
1000
1500
2000
Number of DMIs

Cont.


HINT
A)
P-value is: < 0.001
Observed value is: 91
Frequency of DMI count
0
50
100
150
200
0
20
40
60
80
100
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 600
Frequency of DMI count
0
10
20
30
40
50
60
70
0
100
200
300
400
500
600
Number of DMIs
IntAct
A)
P-value is: < 0.001
Observed value is: 324
Frequency of DMI count
0
50
100
0
50
100
150
200
250
300
350
Number of DMIs
B)
P-value is: < 0.001
Observed value is: 1467
Frequency of DMI count
0
10
20
30
0
500
1000
1500
Number of DMIs

Cont.

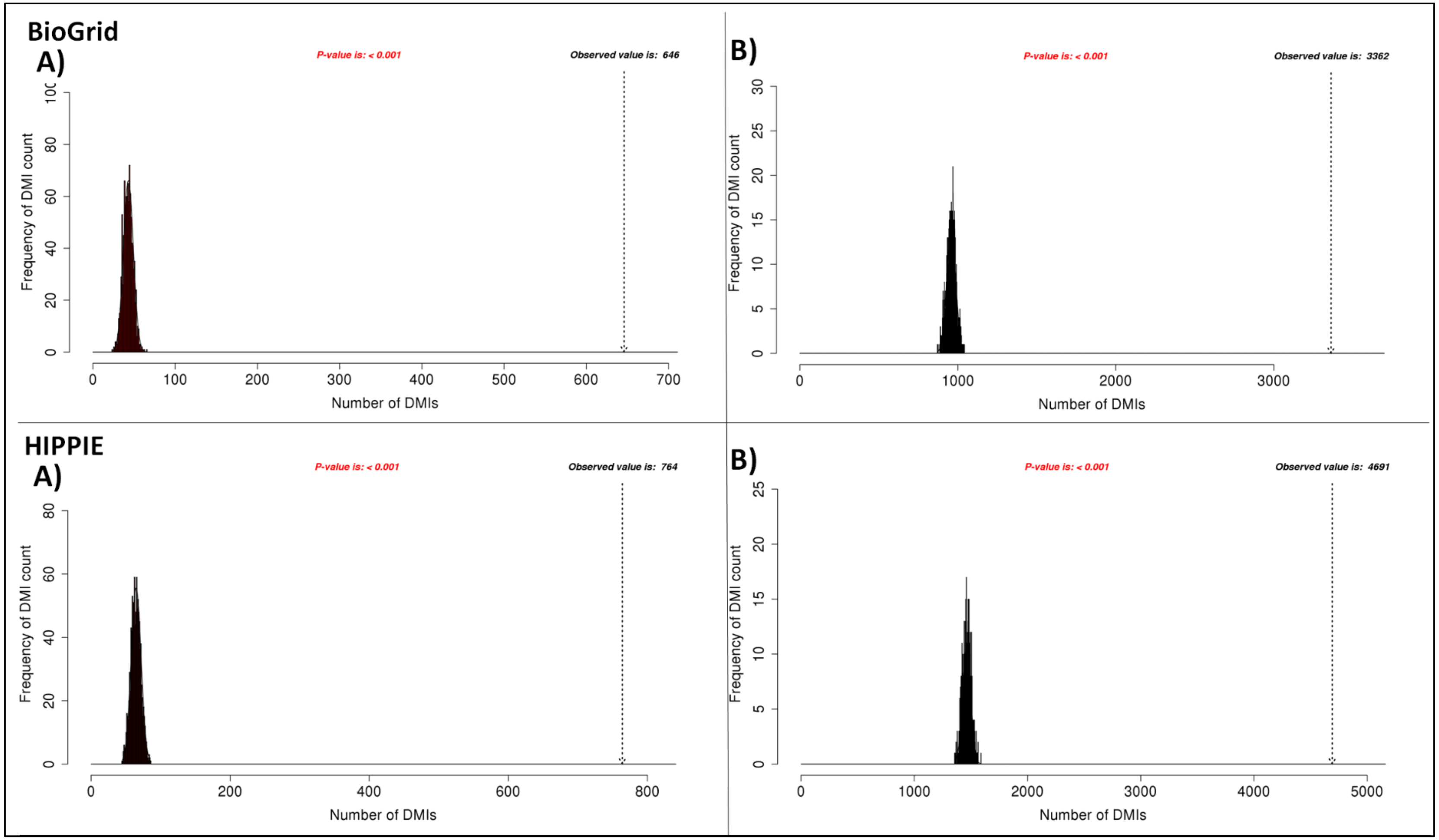


Figure 3.8. Enrichment statistics and histogram of expected random DMI counts in different datasets.

A) Absolute number of DMI count in different datasets using ELMc-Protein strategy, **B)** Absolute number of DMI count in different datasets using ELMc-Domain strategy. Frequency bars indicate the number of randomised PPI datasets returning a given number of known DMIs. The dotted arrow indicates the observed number of known DMIs.

Table 3.4. Comparison of DMI enrichment using different DMI prediction strategies.

DMI prediction strategy	Dataset	Method	potDMIs[1]	DMIs[2]	Enrichment[3] (3 s.f.)	FDR[4] (4 d.p.)
ELMc-Protein	**HI-II-14**	Y2H	162	25	32.1**	0.0311
	BioPlex2.0	AP-MS	318	35	92.3**	0.0108
	QUBIC-15	AP-MS	474	38	14.1**	0.0709
	CoFrac-12	CoFrac-MS	140	4	12.8**	0.0782
	CoFrac-15	CoFrac-MS	223	11	24.0**	0.0417
	HPRD	All	1,485	441	15.8**	0.0633
	HINT	All	1,018	91	10.9**	0.0916
	IntAct	All	1,463	324	12.2**	0.0633
		AP-MS	432	46	8.71**	0.1148
		Y2H	645	48	21.0**	0.0474
	BioGrid	All	1,601	646	15.0**	0.0667
		AP-MS	1,015	129	21.1**	0.0473
		Y2H	964	73	17.0**	0.0586
		CoFrac-MS	409	12	22.2**	0.045
	HIPPIE	All	1,610	764	11.8**	0.0847
		AP-MS	1,518	454	14.8**	0.0387
		Y2H	1,389	240	25.8**	0.0387
		CoFrac-MS	41	2	10.1**	0.0985
ELMc-Domain	**HI-II-14**	Y2H	6,308	74	7.71**	0.1296
	BioPlex2.0	AP-MS	18,105	120	6.87**	0.1456
	QUBIC-15	AP-MS	21,180	164	2.59**	0.3867
	CoFrac-12	CoFrac-MS	4,309	29	2.18**	0.4582
	CoFrac-15	CoFrac-MS	6,479	36	3.95**	0.2532
	HPRD	All	79,663	2,092	4.14**	0.2413
	HINT	All	63,556	600	3.85**	0.26
	IntAct	All	89,824	1,467	2.79**	0.2413
		AP-MS	18,224	178	1.08**	0.9246
		Y2H	21,793	181	4.10**	0.2441
	BioGrid	All	106,370	3,362	3.50**	0.285
		AP-MS	58,305	516	4.02**	0.2487
		Y2H	49,418	305	3.54**	0.2826
		CoFrac-MS	12,676	57	3.05**	0.2598
	HIPPIE	All	105,123	4,691	3.20**	0.3117
		AP-MS	96,947	2,556	2.99**	0.1988
		Y2H	76,681	923	5.03**	0.1988
		CoFrac-MS	704	10	3.02**	0.3311

**P-value < 0.001

[1] Number of all possible DMIs, given the proteins in each dataset.

[2] Predicted SLiM-Protein interactions using known instances.

[3] Observed enrichment of DMIs captured from PPIs

[4] False discovery rate (FDR) calculated as the proportion of the predicted DMIs explained on average by random associations, using the mean random DMI distribution capped at the observed value.

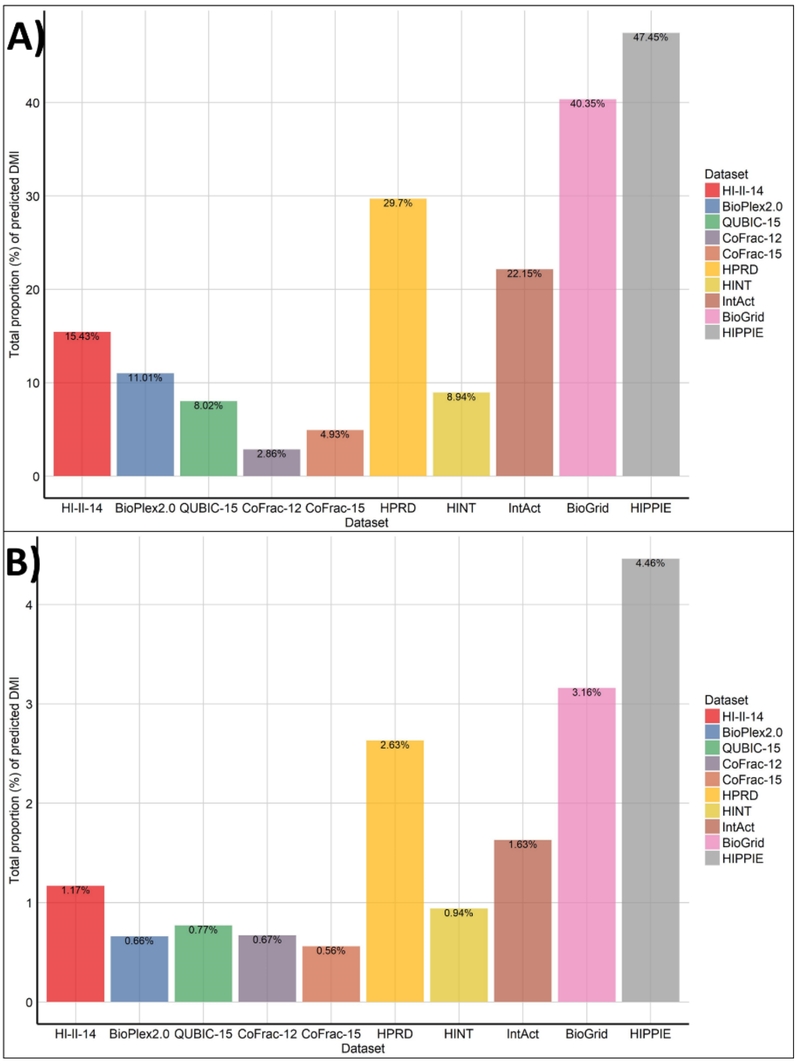


Figure 3.9. Total proportion of DMIs predicted from potential DMIs.

Potential DMIs represents the total proportion of those DMI that were theoretically identifiable given the proteins in the PPIs, Y-axis shows the percentage of predicted DMIs being captured from potential DMIs. **A)** proportion of predicted DMIs captured from potential DMIs using ELMc-Protein strategy, **B)** Total proportion of predicted DMIs captured from potential DMIs using ELMc-Domain strategy.

3.6.5 Impact of PPI quality on enrichment

We then further dig down into how quality of PPIs could influence enrichment. For this purpose, we analysed HIPPIE dataset and evaluated enrichment for PPIs having different confidence scores (0-1) where 1 defines highly confident PPIs. The idea was to see how confidence score was impacting overall DMI enrichment. The PPI quality had impact on the enrichment and it was found that best score to identify DMIs was 0.9 **(Figure 3.10A).** Just like DMI, impact of PPI quality on enrichment was seen, but there was not a simple correlation of PPI confidence with DDI enrichment. Again, the best confidence score for DDI prediction was 0.9 **(Figure 3.10B).**

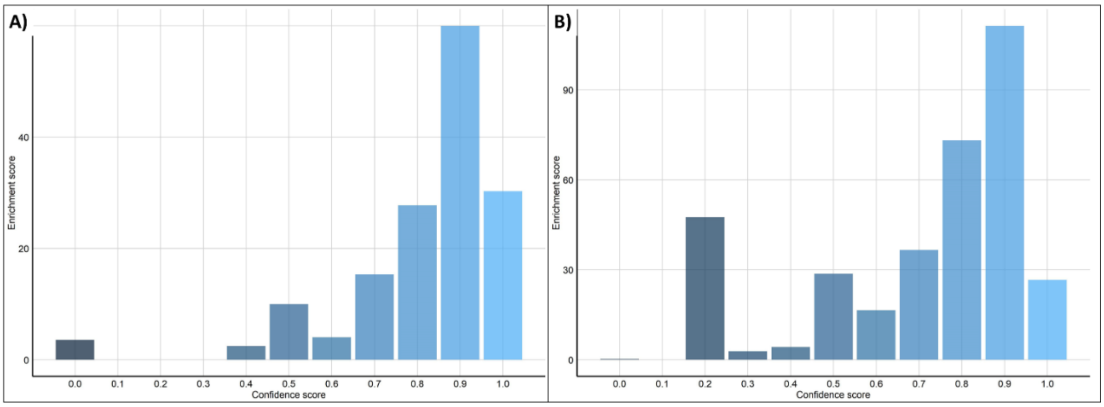


Figure 3.10. Impact of PPI quality on enrichment.

X-axis represents the confidence scores of different PPI subsets from HIPPIE database (0-1). Y-axis represents enrichment score of PPIs belonging to different subsets of PPIs based on confidence scores. A) PPI confidence score vs DMI enrichment score, B) PPI confidence score vs DDI enrichment.

3.6.6 How enrichment changes with random DMI as percentage of total potential DMI

As all databases showed lower enrichment than individual high-throughput screens, it raised the question of whether the apparent lack of enrichment in databases was due to the random expectation being so high. To answer this question, we analysed what proportion of random DMIs was in total potential DMIs.

Only small proportion of random DMIs was in total potential DMIs. Approximately 2-6% random DMIs were being captured from total potential DMIs in BioGrid and HIPPIE datasets. Around 0.5-2.5% random DMIs were being captured from potential DMIs in HPRD and IntAct datasets. All other datasets had ~0-2% random DMIs being captured from total potential DMIs **(Figure 3.11).**

Comparatively, databases except HINT had more random DMIs than individual high-throughput screens. Bioplex2.0 and HI-II-14 had very low proportion of random DMIs, which could be the reason that their enrichment was higher than other datasets. But looking at other high-throughput screens, its clear that they also had fewer random DMIs but their enrichment was not as high as BioPlex2.0 and HI-II-14 **(Figure 3.11).** On the other hand, databases were capturing higher proportion of random DMIs but their enrichment in some cases was close to CoFrac-15. It can be said that the reason of capturing higher proportion of random DMIs could be related to size of the dataset **(Figure 3.11).**

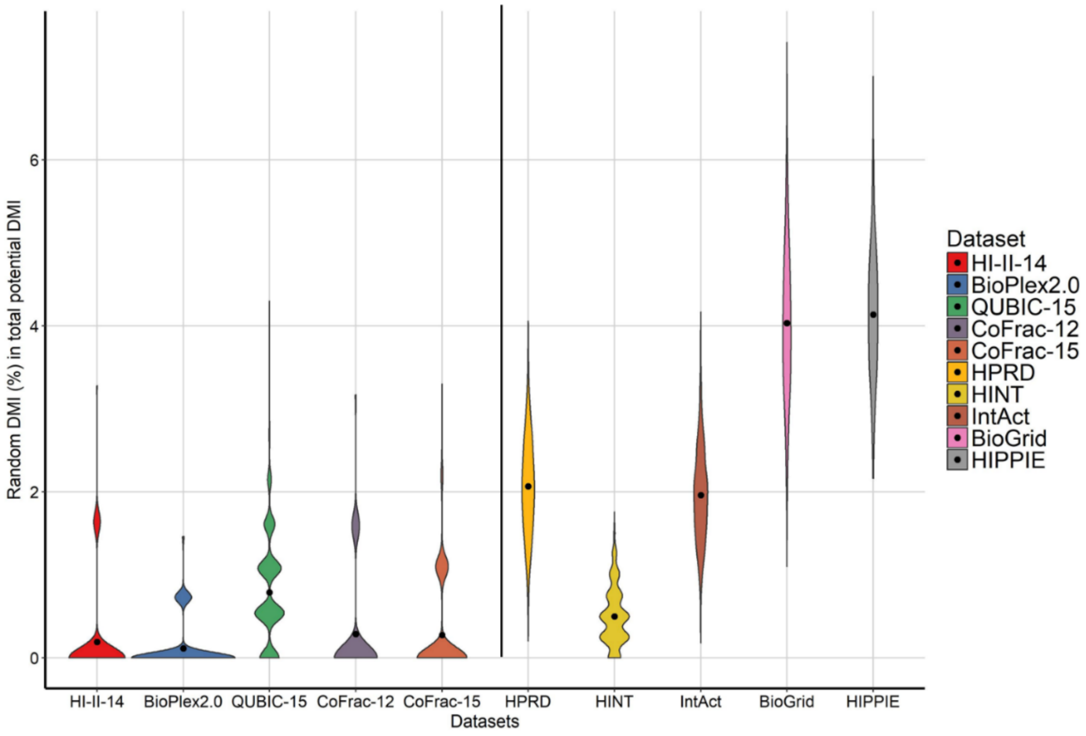


Figure 3.11. Percentage of random DMIs in total potential DMIs.

Y-axis shows the percentage of random DMIs being captured from total potential DMIs. Black dot inside the violin bars is the median of random DMIs. Datasets shown, from left to right, BioGrid, BioPlex2.0, CoFrac-12, CoFrac-15, HI-II-14, HINT, HIPPIE, HPRD, IntAct and QUBIC-15.

We also checked whether quantity of data had any influence on the overall enrichment of DMIs. For this purpose, we compared number of DMIs for each dataset to their respective enrichment scores from ELMi-Protein strategy. We didn't observe any direct impact of DMI numbers on enrichment **(Figure 3.12).**

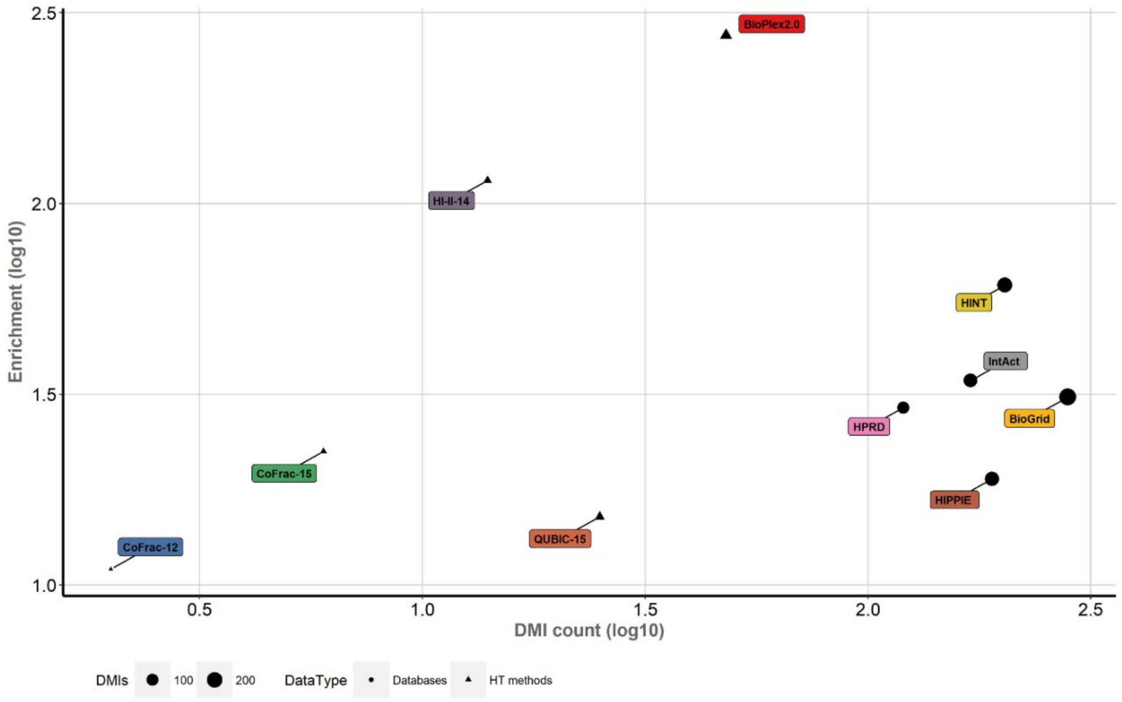


Figure 3.12. Impact of number of DMIs on enrichment.

X-axis is the log10 scale of number of DMIs, y-axis is the log10 scale of enrichment score. Triangle shape is for high-throughput methods, circles are for databases, each point represents single dataset and size of the shape indicates number of DMIs.

3.6.7 Impact of ELM types on capturing DMIs

There are six ELM types known in the ELM database i.e. cleavage (CLV), degron (DEG), docking (DOC), ligand (LIG), post-translational modification sites (MOD) and targeting (TRG). To see if there were any ELM types that were making PPI detection techniques better or worse at capturing DMI, we checked enrichment for individual ELM types using ELMc-Protein. Here, we have combined high-throughput PPI data available in BioGrid, HIPPIE and IntAct databases. PPI pairs were made symmetrical, non-redundant and were restricted to reviewed proteins. All methods captured significant amount of DEG mediated DMIs but didn't capture any significant amount of CLV or TRG mediated DMIs. AP-MS PPIs captured more LIG mediated DMI followed by DEG, DOC and MOD. Two-hybrid captured more DEG mediated DMI followed by LIG, DOC, TRG and MOD. Similarly, CoFrac method captured more DEG mediated DMIs followed by LIG, DOC and MOD **(Figure 3.13).**

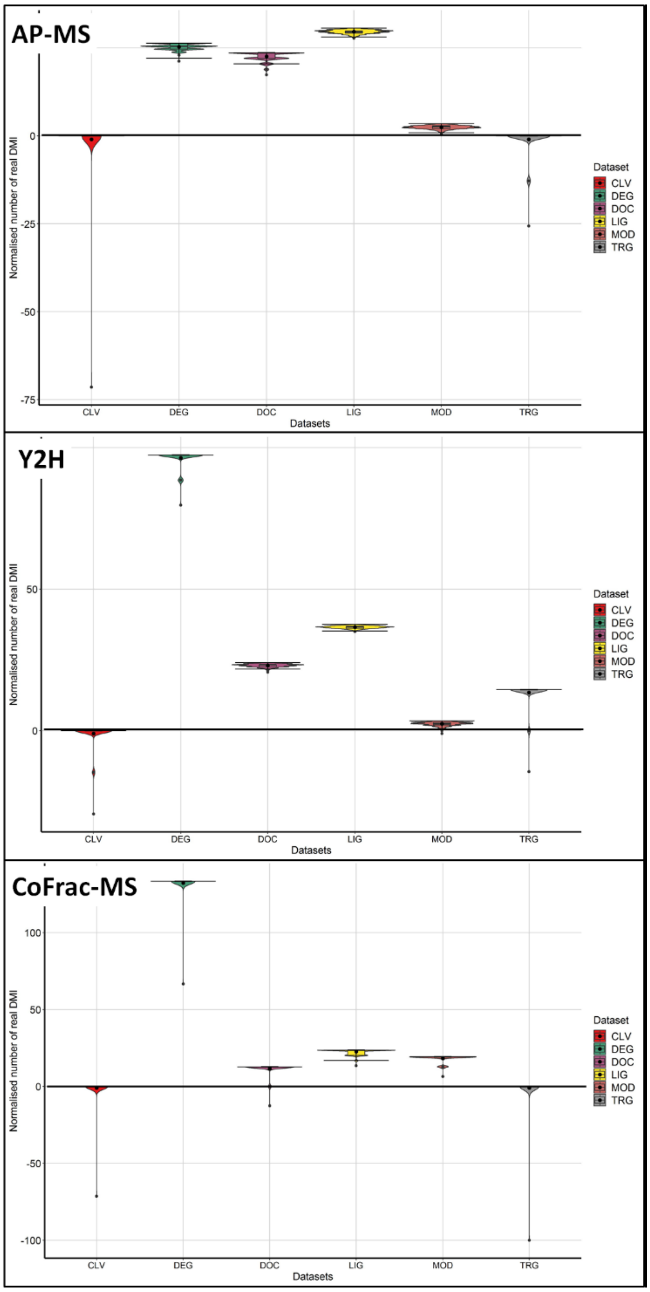


Figure 3.13. Impact of ELM types on DMI enrichment in three high-throughput methods (AP-MS, Y2H and CoFrac-MS).

Top panel shows impact of ELM types in DMIs captured by AP-MS dataset, middle panel shows impact of ELM types in DMIs captured by Y2H dataset and bottom panel shows impact ELM types in DMIs captured by CoFrac dataset.

3.7 Discussion

Different experimental methods have been developed and applied to find domain-motif interactions. Most of these methods have been applied to a limited set of domain families: PDZ, SH2, SH3, and WW which has left many important domain families out of the picture (Blikstad and Ivarsson 2015). Nowadays, studies are being conducted to find SLiMs in conjunction with their binding partners in human proteome (Li, Wu et al. 2010; Zhang, Lin et al. 2015). Most of the available DMI knowledge has been derived from low-throughput studies and there has been no specific study to see how well high-throughput methods capture different sorts of interactions i.e. DMI and DDI. In this study we compared PPIs identified by different groups and databases to see their capability of capturing DMIs and DDIs. One underlying issue in analysing protein interactions is their identifiers. Various research groups and data repositories use different protein identifiers in their analysis which has become a central difficulty. Thus, to be consistent with diverse datasets, it is crucial to map data to a common identifier (Huang, McGarvey et al. 2011). In our study, we have mapped datasets which didn't already had Uniprot IDs (*e.g.* BioGrid, HIPPIE, HPRD, HI-II-14, CoFrac-12 and CoFrac-15) to Uniprot Identifiers for consistency and to avoid any redundancy issues. The number of interactions pre- and post-mapping and filtering looked considerably good, increasing our confidence in prepared datasets for the analysis.

3.7.1 High-throughput screens and public PPI repositories capture DMIs and DDIs

Comparison of data generated by high-throughput screens i.e. BioPlex2.0, HI-II-14, QUBIC-15, CoFrac-12 and CoFrac-15 revealed that all these datasets were significantly enriched in terms of capturing DMIs when compared with random pairs of proteins. The number of DMIs captured by these datasets corresponded to a very low percentage (<1%) of known DMIs available in ELM. On a face value, this might be considered disappointing, but enrichment fold of these datasets suggest that they are indeed capturing real DMIs. Looking at the enrichment trend in different datasets, it was seen that high throughput methods were not notably worse than PPI databases and, in some cases, seem a lot better. BioPlex2.0 and HI-II-14 were the best scorer in terms of capturing DMIs **(Table 3.1, Figure 3.2A).** The general trend of enrichment in datasets suggests that Y2H and AP-MS can both be good methods to study Domain-motif interactions and high-throughput screen of PPIs indeed capture DMIs.

We also checked how different datasets were capturing DDIs. All datasets showed significant enrichment for capturing DDIs when compared with random pairs of proteins. Both BioPlex2.0 and HI-II-14 showed higher enrichment as compared to CoFrac datasets **(Figure 3.2B).** Disappointingly, only a small proportion of PPIs attributed to known DDIs **(Table 3.1).** The total proportion of PPIs that could be explained by DMI or DDI data was quite small. Only a small percentage of PPIs was either being mediated through DMI or DDI **(Figure 3.4).** According to our analysis, the known DMI data captured by different datasets accounts to <1% of the PPIs which is in agreement with previous literature that current number of known DMIs account for less than 1% of interactions (Neduva and Russell 2006). According to Schuster-Bockler and Bateman, the current known DDI data in iPfam can only explain a subset of 4-19% of protein interactions in *Homo sapiens* (Schuster-Bockler and Bateman 2007). In our analysis, the known DDI data from 3did accounts to less than 3% of the PPIs, which again suggests that the high-throughput methods might be depleted in terms of capturing these interactions. This highlights the concern that there is a large proportion of DMIs/DDIs that is yet to be discovered. The number of identified DMIs and DDIs in PPIs and their enrichment suggest that protein composition is important when identifying these interactions. Furthermore, chi-square test showed which datasets had significant enrichment in comparison to other datasets. HPRD was found to be capturing higher proportion of DMIs as compared to other datasets **(Figure 3.4).** Comparison of HPRD (i.e. HPRD > other datasets) with other datasets through chi-square pairwise test showed that HPRD was more significantly enriched than other datasets in terms of capturing DMIs. All other databases had less proportion of PPIs as DMIs than HPRD but higher proportion than high-throughput methods **(Figure 3.4).** IntAct was capturing significantly more DMIs as compared to HIPPIE, HINT, BioGrid, HI-II-14, CoFrac-12, CoFrac-15 and BioPlex2.0. High-throughput methods on the other hand had lower proportion of DMIs in comparison to curated databases **(Figure 3.4).** Curated databases were generally found significant in comparison to other datasets **(Table 3.2, Figure 3.4)**.

HPRD showed similar trend in case of DDIs. CoFrac-12 showed significant difference with IntAct, HIPPIE, HINT, BioGrid and QUBIC-15. BioPlex2.0 showed significant difference with HIPPIE. HI-II-14 showed significant difference with IntAct. CoFrac-15 showed significant difference with HIPPIE, IntAct, HINT and BioGrid. QUBIC-15 showed significant difference with IntAct. Overall, high-throughput methods looked better in terms of having higher proportion of DDIs and were more significantly enriched in terms of capturing DDIs as compared to databases except HPRD which was found to be significantly enriched than any other dataset **(Table 3.2, Figure 3.4)**.

In future, it would be interesting to have a comparative study where different other methods including BioID Mass Spectrometry (Li, Meng et al. 2019) and Phage Display (Sidhu, Fairbrother et al. 2003) can be compared to see if they are any good at capturing DMIs/DDIs.

3.7.2 Binary vs Co-complex mapping

The two orthogonal methods of mapping PPIs are binary where two proteins are in direct physical contact with each other and co-complex where interactions usually require additional proteins to form multimeric complexes. These complexes can have both direct and indirect interactions of different proteins. Y2H is widely known to identify binary interactions whereas AP-MS and CoFrac-MS are being used to identify co-complex interactions (Luck, Sheynkman et al. 2017). As both HI-II-14 and BioPlex2.0 datasets showed higher enrichment fold (~120x), and CoFrac datasets didn't perform well, we decided to further find out which method (i.e. binary and co-complex) was better in terms of identifying DMIs and DDIs. We extracted binary PPIs (i.e. Y2H) and co-complex PPIs (i.e. AP-MS and CoFrac-MS) from three well known databases (i.e. BioGrid, IntAct and HIPPIE) to evaluate enrichment. In previous studies e.g. (Hecker, Rabiller et al. 2006; Hu, Song et al. 2009), binary approaches have been used to identify DMIs for example SUMO interacting motifs have been identified that interact with the SUMO1 and SUMO2 proteins (Hecker, Rabiller et al. 2006; Hu, Song et al. 2009) though there are no specific studies where co-complex approaches have been used to discover DMIs. In our analysis, all datasets showed significant DMI enrichment when compared with random pairs of proteins and there was no clear winner among Y2H and AP-MS. They both were capturing significant number of known DMIs. CoFrac-MS on the other hand didn't perform well **(Table 3.3, Figure 3.2A).** Y2H and CoFrac data significantly captured degron motif mediated DMIs while AP-MS captured more conventional ligand mediated DMIs **(Figure 3.13).** ELM type analysis revealed that CLV and MOD were not generally good at capturing DMIs. The reason that other ELM types captured significant DMIs and CLV and MOD didn't could be motif complexity or their low complexity nature and involvement in post-translational modifications. As supported by the ELM type analysis, CoFrac data is not so good as it is only getting complexes but still is enriched for DMIs, indicating that DMI are playing important roles in complexes, and should not be thought of only in terms of binary interactions. Overall, ELM type analysis revealed that MOD and CLV types were not generally good at capturing DMIs **(Figure 3.13).** The general trend of enrichment was same in DDIs where both Y2H and AP-MS looked better than CoFrac-MS PPIs **(Table 3.3, Figure 3.2B).**

3.7.3 Overlap between the datasets

As we know, most of the PPI data generated by different studies is also available in public PPI databases therefore, there is a likelihood that these databases share certain overlap of data. As BioGrid and IntAct are known as the most comprehensive PPI databases having most of the high-throughput data available in them. It was likely that DMIs captured by other datasets would also be in them. Therefore, we checked what percentage of DMIs captured by different high-throughput studies and databases was overlapping with each other**.** As expected, most of the DMIs captured by different high-throughput methods were overlapping with each other **(Figure 3.6A)**.

3.7.4 DMI prediction quality impacts enrichment

As it can be seen, the big PPI databases have large number of PPIs, of which only few are known to be mediated by DMIs. Looking at the enrichment of these databases, it was clear that these big databases are capped on enrichment due to the low number of known DMIs and there are many more DMIs that are yet to be found. To see if adding noise in DMI prediction can help discover more DMIs with significant enrichment, we tested different DMI prediction methods. During recent years, several studies have combined PPI data with computational approaches to identify new DMI for known recognition domains, SH2, SH3, PDZ and WW domains (*e.g.* (Encinar, Fernandez-Ballester et al. 2009; Pichlmair, Kandasamy et al. 2012; Weatheritt, Jehl et al. 2012; de Chassey, Meyniel-Schicklin et al. 2014), we combined PPI data with ELM and domain information available in ELM database as DMI prediction strategies. We found that with the introduction of noise in DMI network (using ELM and/or domain information), number of DMIs increased while the overall enrichment score dropped for all datasets **(Table 3.4).** In general, the overall enrichment trend/ranking of datasets almost remained same when we used ELMc-Protein **(Figure 3.7A)** or ELMc-Domain **(Figure 3.7A)** strategies. This was in agreement with our previous analysis that noise in DMI network lowers enrichment score **(Figure 2.4)** (Idrees, Perez-Bercoff et al. 2018). The predicted DMIs from ELMc-Protein strategy had a very low false discovery rate (FDR) showing their possibility of being real. On the other hand, the estimated FDR for individual DMI predictions was quite high (0.1-0.9) for ELMc-Domain strategy which highlights the need for caution when interpreting naïve large-scale predictions of this nature **(Table 3.4)**. Overall, it can be seen from all sorts of DMI predictions that all these databases were quite enriched in terms of capturing DMIs**.** Motif predictions from different tools (for example, SLiMProb) with conservation masking can result in less noisy predictions and in future, it would be interesting to investigate it in depth to see how it impacts quality of DMI predictions.

3.7.5 PPI confidence does not equate to quality

HINT and HPRD were our high confidence datasets. Both datasets were quite enriched in terms of DMIs as well as DDIs. Both databases showed higher DMI enrichment than IntAct and HIPPIE but had lower enrichment than BioGrid. On the other hand, HPRD was the most enriched dataset in terms of capturing DDIs while HINT didn't perform as well. Among databases, HINT was better than BioGrid but had lower enrichment than other databases. HI-II-14, BioPlex2.0 and CoFrac-15 were more enriched than databases in terms of capturing DMIs. HI-II-14, BioPlex2.0 and QUBIC-15 had higher DDI enrichment than all databases except HPRD. The two CoFrac-MS datasets were the lowest DDI enriched datasets. To see if PPI quality had anything to do with enrichment, we evaluated enrichment in PPIs in HIPPIE database based on their confidence score. Looking the enrichment trend, we didn't observe any direct impact of confidence score on enrichment **(Figure 3.10)** which means confidence does not necessarily equate to quality of data for DMI or DDI predictions.

3.8 Conclusion

High-throughput experimental methods are generating large number of protein-protein interaction (PPI) data. New methods for assessing the quality of identified PPIs as a source of different types of interactions (*i.e* Domain-Motif Interaction (DMI) or Domain-Domain Interaction (DDI)) are in high demand due to the error prone nature of these methods. In our current analysis, we have assessed PPIs identified from different high-throughput screens and publicly available databases as a source of capturing DMIs and DDIs. We found significant enrichment in all databases, both Y2H and AP-MS looked promising in terms of capturing DMIs and DDIs whereas CoFrac might not be a good source of capturing these interactions.

4 Chapter 4: Bioinformatics prediction of molecular mimicry in viruses

4.1 Abstract

Viruses cause dreadful diseases in humans through establishing protein-protein interactions (PPIs) with the host cells. During recent years, exponential growth has been seen in our knowledge of viral interaction networks. Viruses hijack host cellular machinery through mimicking short linear motifs (SLiMs) in host proteins to maintain their life cycle inside host cells. Although the number of vhPPIs has grown over the years in databases (*i.e*. PHISTO and VirHostNet2.0), the prediction of molecular mimicry is still considered challenging because of their degenerate nature of SLiMs. For this reason, new computational methods are much needed to predict new mimicry instances in viruses. In this chapter, the SLiMEnrich computational pipeline developed in Chapter 2 is applied to study molecular mimicry by viruses using public vhPPI data. The result of this chapter shows that vhPPIs available in the PHISTO and VirHostNet2.0 databases capture domain-motif interactions (DMIs). In chapter 3, I found that both AP-MS and Y2H were good methods to capture DMIs in humans. Analysis in this chapter agrees with these findings that vhPPIs identified through Y2H and AP-MS are capable of capturing DMIs. Y2H captured more DMIs (8 in case of PHISTO and 6 in case of VirHostNet2.0) as compared to the AP-MS (6 in case of PHISTO and 4 in case of VirHostNet2.0). The FDR of captured DMIs was quite low where FDR of DMIs captured by Y2H was (0.0338-0.0366) while the FDR of DMIs captured by AP-MS was (0.0253-0.0860). Comparison of viral subtypes revealed that dsRNA viruses were more enriched than ssRNA viruses in terms of DMIs within the available PPI data. On the other hand, ssDNA showed more enrichment than dsDNA viruses. If we compare RNA vs DNA viral interactions, RNA viruses were more enriched for DMIs than DNA viruses. The derived knowledge from this Chapter and Chapter 3 was used to predict novel SLiMs using a *de novo* SLiM discovery tool, QSLiMFinder. A total of 2,316 motifs were

predicted in 1,715 significant datasets. In conclusion, it can be said that vhPPI data can be used to discover new DMIs and SLiMs.

4.2 Introduction

Viruses are obligate parasites that replicate inside host cells by establishing interactions with host proteins (Benedict, Norris et al. 2002; Finlay and McFadden 2006). Viruses are responsible for life-threatening diseases in humans. To prevent and treat these diseases, it is crucial to understand host-pathogen biological systems (Jean Beltran, Federspiel et al. 2017). Virus-host protein-protein interactions (vhPPIs) are a regular event that occur throughout the viral life cycle. Viruses replicate inside host cells through hijacking host cellular machinery (i.e. proteins, lipids and metabolites), which is often achieved by "molecular mimicry" of host interactions (Neduva and Russell 2005; Davey, Trave et al. 2011; Chaurushiya, Lilley et al. 2012; Davey, Van Roey et al. 2012).

Studying molecular mimicry has become one of the most intriguing aspects of viral research. The term 'molecular mimicry' was first referred as sharing of antigens between pathogens and hosts (Damian 1964), also known as 'antigenic mimicry' (Kohm, Fuller et al. 2003). The classic definition of molecular mimicry can be defined as: the sharing of short stretches of linear amino acid sequence or conformational fit between pathogen and host (Oldstone 1998). Molecular motif mimicry gives advantage to the viruses to replicate and colonize effectively in their host cells (Davey, Trave et al. 2011; Chaurushiya, Lilley et al. 2012) as well as to escape detection during invading the host cells (Benedict, Norris et al. 2002; Finlay and McFadden 2006). In this chapter, the focus is on PPI motif mimicry and how it helps viruses to hijack host cellular machinery.

SLiMs in pathogenic viral proteins are known as mimicry motifs as they have similar, if not identical, amino acid sequences and functions as host SLiMs. SLiMs are robust and highly evolvable elements in viruses, which can lead to rewiring of the vhPPIs (Neduva and Russell 2005; Davey, Van Roey et al. 2012; Chemes, de Prat-Gay et al. 2015). In most of the cases, a SLiM that is adequately exposed on protein surface can regulate biological pathways by affecting protein stability, ligand binding and targeting (Neduva and Russell 2005; Van

Roey, Uyar et al. 2014). Various examples of mimicry motifs have been reported in different pathogens, especially in proteins involved in attachment, penetration and cytoadherence. One of the best-known examples in viruses is the polyproline motif (PxxPxR), which has been reported in non-structural 5A protein (NS5A) of hepatitis C virus as well as in Nef protein of HIV type 1 (Shelton and Harris 2008). This polyproline motif establishes interactions with SH3 domains of the host proteins (Shelton and Harris 2008). Another widely known example of motif mimicry is by human papilloma virus E7 protein which mimics LxCxE motif and disrupts the functionality of tumour suppressor retinoblastoma protein 1 in host cells **(Figure 4.1)** (Chemes, de Prat-Gay et al. 2015)**.**

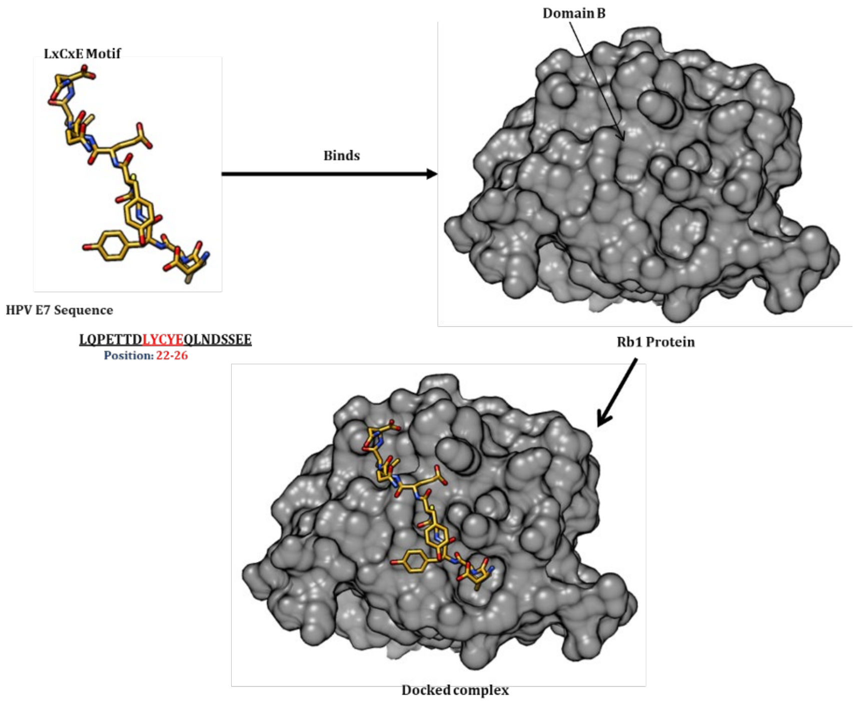


Figure 4.1. A known example of motif mimicry.

B domain of retinoblastoma (Rb1) tumour suppressor protein is inactivated by binding of LxCxE motif of human papilloma virus E7 protein. The 3D resolved structure of the Rb protein shows a linear peptide containing LxCxE motif sequence (position: 22-26) bound to a highly conserved groove on Rb_B domain of Rb1 protein (PDB ID: 1GUX) (Lee, Russo et al. 1998).

High-throughput methods (*i.e.* Y2H and AP-MS) have been applied to understand underlying basis of vhPPIs for example HCV proteome mapping has provided information on the molecular basis of co-deregulation of insulin and TGF-β signalling pathways (de Chassey, Navratil et al. 2008; Hagai, Azia et al. 2011; de Chassey, Meyniel-Schicklin et al. 2014). Y2H has been regarded as the most popular method to resolve viral-human interactomes while AP-MS approach has been used to map interactomes of only 30 viral species (Pichlmair, Kandasamy et al. 2012; de Chassey, Meyniel-Schicklin et al. 2014). Both these methods have their strengths and weaknesses. Y2H is good at identifying binary biophysical interactions but is not compatible with self-activating or membrane proteins. Y2H interactions are generally not in their native biological context; to get high confidence vhPPIs, several factors need to be considered (i.e. yeast strains, reporter genes, plasmid copy number, stringency conditions, and fusion proteins). On the other hand, AP-MS is considered better in identifying context-dependent vhPPIs and can be done under more physiological conditions **(Chapter 1: 1.3)**. During recent years, different computational methods have been developed to study proteome wide vhPPIs (Dyer, Murali et al. 2007; Evans, Dampier et al. 2009; Segura-Cabrera, Garcia-Perez et al. 2013). But most of these studies have been targeted to selected pathogens only (Emamjomeh, Goliaei et al. 2014; Barnes, Karimloo et al. 2016; Zhang, He et al. 2017). To date, there has been no study to analyse different viral subtypes to see how they perturb host cellular machinery for their regulatory functions and infection cycle through hijacking SLiMs. Therefore, it's of interest to see how different subtypes of viruses tend to interact with host proteins through SLiMs. Thus, this chapter is focused on utilization SLiMEnrich to study mimicry using vhPPI data, to see which high-throughput method is better at predicting mimicry, to see how different viruses hijack host cellular machinery and to discover new SLiMs.

4.3 Aims and objectives

The main objective was to combine the results of Chapters 2 and 3 with available vhPPI data to gain insight into SLiM-mediated interactions between viruses and their hosts, with a specific focus on motif mimicry. More specifically, this chapter aims to:

- Assess whether the available vhPPI data is enriched for known SLiM-mediated interactions and identify which large-scale PPI capturing method (two hybrid and affinity purification) is better for studying SLiM-mediated interactions in virus-host context.
- Analyse viral subtypes to infer how much they use SLiMs to perturb host cellular machinery for their regulatory functions and infection cycle.
- Use derived knowledge for *de novo* prediction of novel SLiMs involved in viral molecular mimicry.

4.4 Methods

4.4.1 Data acquisition and enrichment analysis

Two comprehensive virus-host PPI (vhPPI) databases were downloaded: Pathogen Host Interaction Search Tool (PHISTO) (Durmus Tekir, Cakir et al. 2013) [retrieved on: 2018-05-24] and Virus Host Network 2.0 (VirHostNet2.0) (Guirimand, Delmotte et al. 2015) [retrieved on: 2018-09-11]. vhPPIs belonging to four viral groups (ssRNA, dsRNA, ssDNA and dsDNA) were individually downloaded from VirHostNet2.0 database as well as PHISTO database. Both databases were split into two well-known high throughput methods, Y2H and AP-MS, by pulling out interactions using "two hybrid" and "affinity" as keywords. ELM data from our previous analysis **(Chapter 3: 3.2.1)** was used to evaluate enrichment and to predict DMIs using vhPPI data. Enrichment differences were evaluated using SLiMEnrich **(Chapter 2)** through the ELMi-Protein strategy **(Figure 2.1).** ELMc-Protein and ELMc-Domain strategies **(Figure 2.1)** were employed to further increase the size of the network and to discover new DMIs. Normalisation of the data was done by dividing number of real DMIs by mean random DMIs.

4.4.2 DMI prediction using predicted SLiM instances

To predict new DMIs, new SLiM instances of known ELMs were predicted using SLiMProb v2.5.1 (Edwards and Palopoli 2015) with the disordered masking feature (IUPred score >= 0.2) (Hagai, Azia et al. 2011). The predicted SLiMs were then used to predict DMIs using SLiMEnrich **(Chapter 2)** through the ELMc-Protein (predicted SLiMs mapped to known human partner proteins via ELMs) and ELMc-Domain (predicted SLiMs mapped to Pfam-domain-containing human partner proteins) strategies **(Figure 2.1)**.

4.4.3 DMI prediction in different viral subtypes

VirHostNet2.0 and PHISTO datasets were divided into four viral groups based on their genetic material. PHISTO dataset had viral PPIs from 8 different double stranded DNA virus families (dsDNA), 2 single stranded DNA virus families (ssDNA), 2 double stranded RNA virus families (dsRNA) and 13 single stranded RNA virus families (ssRNA). VirHostNet2.0 data consisted of 25 ssRNA virus families, 13 dsDNA virus families, 5 ssDNA virus families and 2 dsRNA families. The ELMc-Domain strategy of SLiMEnrich **(Figure 2.1)** was employed to predict DMIs in each group using SLiMs predicted using SLiMProb v2.5.1. Modification (MOD) and cleavage (CLV) ELM types were excluded from the analysis to reduce noise in the network, because post translational modification motifs tend to have lower complexity (and therefore more random occurrences) and interact with common domains (and therefore more chance for random motif-domain connections) **(Chapter 3: 3.6.7).**

Gene ontology (GO) enrichment analysis of human genes targeted by viral proteins was done using Biological Networks Gene Ontology tool (BiNGO) (Maere, Heymans et al. 2005). A binomial statistical test was applied to visualise overrepresentation of the GO terms (p-value < 0.001). Benjamini & Hochberg false discovery rate correction was applied for testing correction. Human proteins from the vhPPI were selected as the background to evaluate enrichment of DMI (mimicry) targets over general viral targets.

4.4.4 *De-novo* prediction of human SLiMs mimicked by viruses

As both Y2H and AP-MS datasets showed significant DMI enrichment therefore, I decided to focus on one of them (HI-II-14) to further explore them and to see if PPIs having significant DMI enrichment could be used for *de-novo* SLiM predictions. The reason of selecting HI-II-14 dataset was two-fold: 1) It was found to be significantly enriched for real DMIs/DDIs, 2) It was based on Y2H experiment which has previously be shown to be effective in terms of capturing DMIs (Blikstad and Ivarsson 2015).

To do this, I integrated viral (PHISTO) (Durmus Tekir, Cakir et al. 2013) and human (HI-II-14) (Rolland, Tasan et al. 2014) datasets together by mapping protein partners of each viral protein in vhPPIs to their respective interactors in human interactome.

A total of 12,139 datasets were generated for vhPPI pairs; for each pair the dataset contained a single viral protein and all the human interactors of the viral protein's human interaction partner. FASTA sequences for each dataset were retrieved from the Uniprot database (UniProt Consortium 2018) and were fed to QSLiMFinder v2.20 (Palopoli, Lythgow et al. 2015), [ambiguity=T and cloudfix=F] with the viral protein in each dataset treated as the query sequence for *de-novo* discovery of SLiMs. QSLiMFinder looks for any sequence motifs present in this query sequence that are enriched in the rest of the dataset (e.g. viral protein motifs that are enriched in the human interaction partner).

Two alternative versions of the integrated dataset were simulated as control groups:

1. Random viral protein ("randomvProtein"): The vhPPI network was disrupted by shuffling viral proteins. This pairs each viral protein with the human interactors of a random human protein.
2. Random human interactor ("randomInteractor"): The human-human PPI network was disrupted by shuffling human proteins. This effectively pairs each viral protein with a random set of human proteins.

As with the real data, FASTA sequences of all proteins were retrieved from Uniprot (UniProt Consortium 2018) and used for the *de-novo* discovery of SLiMs using QSLiMFinder v2.20 (Palopoli, Lythgow et al. 2015) with the viral proteins used as the query **(Figure 4.2).** The P-value of each SLiM returned was estimated using default QSLiMFinder "Sig" values.

4.4.5 Multiple testing correction for de novo SLiM prediction

Multiple testing correction for the QSLiMFinder predictions was performed by calculating the estimated approximate FDR based on the expected number of false positives, using:

$$\mathrm{FDR} = p\mathrm{N}/n_p,$$

Where N represents the total number of datasets, n_p represents number of results returned with significance p-value. Note that, unlike a traditional statistical test, a single dataset might return multiple true and/or false positive SLiM predictions. Datasets that were too small (too few UPC) were disregarded from the analysis. As per QSLiMFinder default settings (Palopoli, Lythgow et al. 2015), only datasets that had 3+ unrelated proteins (UPCs) were included in the analysis and analysis was focused on significant datasets (QSLiMFinder default, p-value <0.1).

4.4.6 Comparison with previously published ELMs

CompariMotif v3.13 (Edwards, Davey et al. 2008) was used to compare discovered motifs with the previously published motifs from ELM and to find degree of overlap and relationships between them. Motifs were then classified using the benchmarking criteria from the QSLiMFinder paper (Palopoli, Lythgow et al. 2015): a motif was regarded as a true positive (TP) match if it met minimum match criteria of MatchIC ≥ 1.5 and normalised IC ≥ 0.5, and the hub protein was known to interact with the matching ELM; a motif was regarded as off-target (OT), if the pattern matched an ELM with more stringency (MatchIC ≥ 2.5 or NormIC ≥ 1.0) but the matched ELM was not known to interact with the hub protein. Any hits below the minimum match criteria were regarded as spurious and ignored. Motifs without any matches meeting the criteria were considered false positive (FP) predictions if returned by control datasets, or candidate novel motifs if returned by the real data.

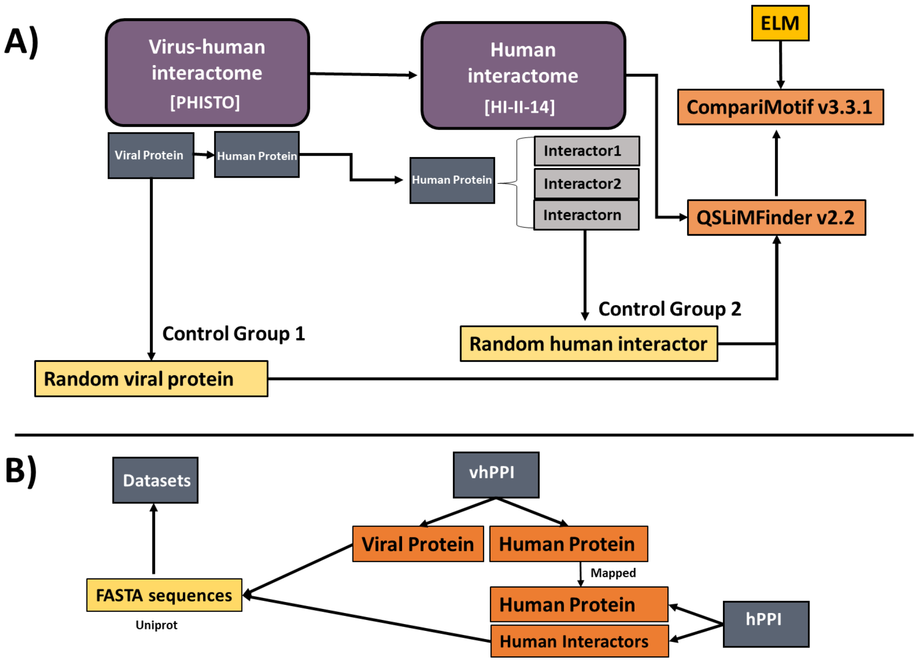


Figure 4.2. Schema of *de-novo* SLiM discovery and data generation.

A) Basic workflow of *de-novo* SLiM discovery pipeline. The virus-human interactome is integrated with the human interactome by mapping each human partner protein (hProtein) to its corresponding human interaction partners (Interactors) in HI-II-14 dataset. The human interactors were then added to the corresponding viral protein (vProtein) to make a dataset for QSLiMFinder v2.2 de-novo discovery of SLiMs, using the vProtein as the query. Two control groups were generated where the first group had shuffled viral proteins and second group had shuffled human interactor proteins. Discovered SLiMs were then compared with previously published SLiMs from ELM using CompariMotif v3.3.1 tool.

B) Dataset generation for *de-novo* SLiM discovery. vhPPIs and hPPIs are integrated by mapping human protein partners of viral proteins in vhPPIs to human proteins in hPPIs.

4.5 Results

4.5.1 Does the viral-human PPI data capture SLiM based interactions?

In this chapter, I analyse domain-motif interactions (DMIs) in which a SLiM-containing viral protein interacts with an ELM-binding human protein, using virus-host protein-protein interaction (vhPPI) data from the PHISTO (Durmus Tekir, Cakir et al. 2013) and VirHostNet2.0 (Guirimand, Delmotte et al. 2015) databases.

First, it was of interest to assess whether these vhPPI data capture DMIs, and how enriched different datasets are in terms of capturing DMIs. For this purpose, the ELMi-Protein strategy of SLiMEnrich **(Chapter 2, Figure 2.1)** was used**.** VirHostNet2.0 captured 16 known DMIs, which is 20x enrichment compared to random **(Figure 4.3)**. PHISTO captured 22 known DMIs with 18x enrichment **(Figure 4.3)**. Enrichment in both datasets was strongly significant (P-value <0.001)**.** This showed that vhPPI data was indeed capturing DMIs and thus, can be a good source of studying molecular mimicry in viruses.

4.5.2. Which high-throughput method is better at capturing viral-host DMIs?

In the previous analysis of human interactomes, it was found that both Y2H and AP-MS were potentially good methods to capture DMIs **(Chapter 3)**. Keeping that in mind, I evaluated vhPPI data sources to see if they agree with my previous findings. Y2H and AP-MS interactions were extracted from PHISTO and VirHostNet2.0 database and were evaluated for enrichment using ELMi-Protein strategy **(Table 4.1)**. As the number of DMIs was quite small, the noisier ELMc-Protein strategy was also employed; Chapter 3 highlighted ELMc-Protein as the most effective strategy as it captured reasonable number of DMIs with significant enrichment **(Chapter 3: 3.7.3)**. Both strategies use known SLiM instances to assess how well these high-throughput methods are capturing DMIs. Both methods showed significant enrichment (P-value < 0.001) in terms of capturing DMIs **(Table 4.1, Figure 4.3)**. AP-MS PPIs in PHISTO showed higher enrichment as compared to Y2H while AP-MS PPIs in VirHostNet2.0 showed slightly lower enrichment than Y2H **(Figure 4.3)**. The results agreed with our previous analysis that both Y2H and AP-MS are potentially good to captured DMIs.

It should be noted that the FDR shown in these tables is the FDR calculated by SLiMEnrich for each DMI rather than the multiple testing correction of the p-value.

Table 4.1. DMI enrichment in high-throughput interaction data available in PHISTO and VirHostNet2.0 databases.

Strategy	Dataset	Method	vhPPI[1]	potDMI	DMI[2]	Enrichment (3 s.f.)	FDR (4 d.p.)
ELMi-Protein	PHISTO	All	34,832	39	22**	18.0	0.0261
		Affinity	9,973	10	6**	39.5	0.0253
		Y2H	7,701	21	8**	29.5	0.0338
	VirHostNet2.0	All	22,886	30	16**	20.0	0.0498
		Affinity	5,765	7	4**	11.6	0.0860
		Y2H	8,530	19	6**	27.3	0.0366
ELMc-Protein	PHISTO	All	34,832	150	36**	19.0	0.0523
		Affinity	9,973	28	8**	22.6	0.0442
		Y2H	7,701	54	14**	17.5	0.0572
	VirHostNet2.0	All	22,886	101	34**	18.2	0.0547
		Affinity	5,765	31	17**	16.4	0.0908
		Y2H	8,530	52	12**	17.4	0.0575

** P-value < 0.001

1. Non-redundant vhPPIs.
2. Non-redundant observed DMI.

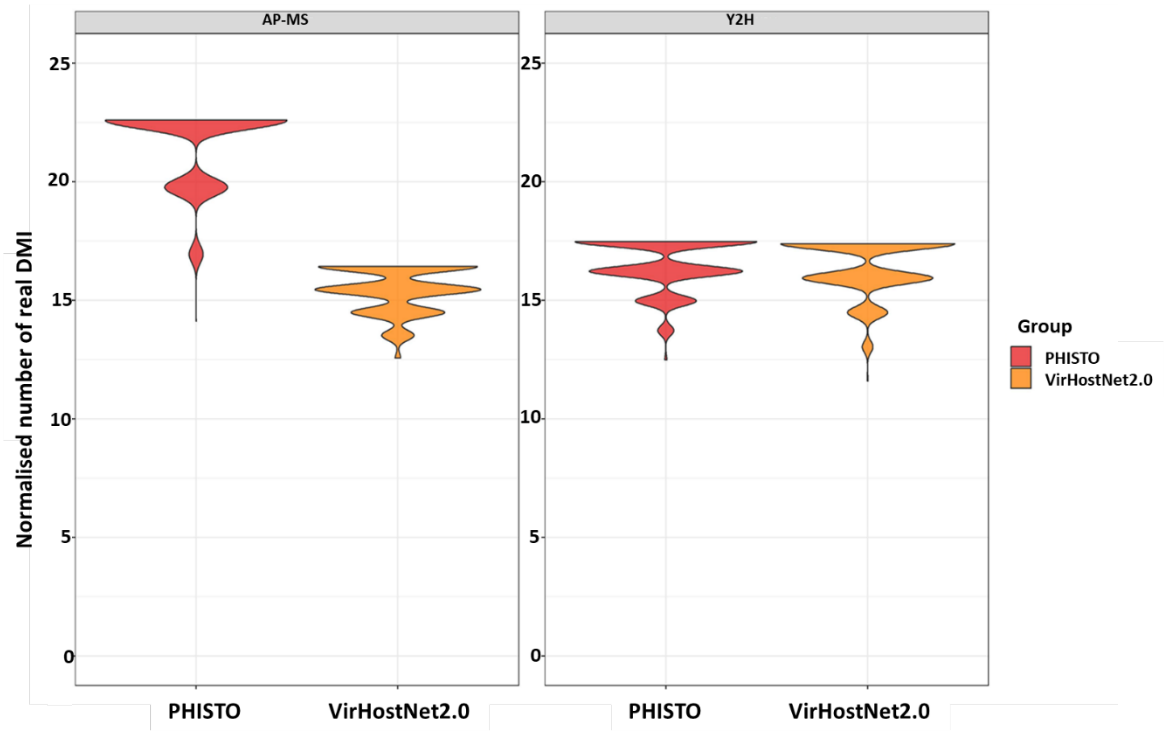


Figure 4.3. Normalised number of real DMIs (DMI $_{Real}$ = DMI $_{Obs}$ – DMI $_{Ran}$).

Real DMIs captured by the high-throughput methods available in PHISTO and VirHostNet2.0 over 1000x randomisations using ELMc-Protein strategy. Y-axis shows the normalised number of real DMIs. Left panel, enrichment of DMIs captured by AP-MS in both datasets. Right panel, enrichment of DMIs captured by Y2H in both datasets. Red, PHISTO. Orange, VirHostNet2.0.

Once it was established that vhPPI data was capturing DMIs and both Y2H and AP-MS methods were good in terms of capturing DMIs, I shifted focus towards predicting new DMIs. For this purpose, all vhPPIs available in PHISTO and VirHostNet2.0 were used.

4.5.3. DMI prediction using known viral instances

For this analysis, I used known viral instances that were known for mimicry in ELM. The ELMc-Protein strategy was used to link known viral instances to their potential human partners known to interact with that ELM class. PHISTO captured 36 non-redundant DMIs with enrichment score of 19.1 and FDR of 0.0523. On the other hand, VirHostNet2.0 captured 35 DMIs (34 non-redundant DMIs) with enrichment score of 18.2 and FDR of 0.0547 **(Figure 4.6B)**. 25 DMIs were captured by both datasets **(Figure 4.4, Table 4.2).**

Table 4.2. Known/Predicted DMIs captured by PHISTO and VirHostNet2.0 datasets.

Dataset	**vProtein**	**Uniprot**	**Motif**	**hProtein**	**Uniprot**
PHISTO	LT-Ag	B8ZX42	LIG_Rb_LxCxE_1	RB1	P06400‡
	E7	**P03129**	**LIG_Rb_LxCxE_1**	**RB1**	**P06400 ‡**
	LMP1	**P03230**	**LIG_TRAF2_2**	**TRAF2**	**Q12933 ‡**
	UL48	P06492	LIG_HCF-1_HBM_1	HCFC1	P51610
	LMP2	**P13285**	**LIG_WW_1**	**NEDD4**	**P46934**
	LMP2	P13285	LIG_WW_1	ITCH	Q96J02
	UL56	**P28282**	**LIG_WW_1**	**NEDD4**	**P46934‡**
	UL56	P28282	LIG_WW_1	ITCH	Q96J02
	Segment-10	**P08363**	**LIG_WW_1**	**NEDD4**	**P46934‡**
	Segment-10	P08363	LIG_WW_1	ITCH	Q96J02‡
	VP-40	Q05128	LIG_WW_1	NEDD4	P46934
	EBNA6	**P03204**	**LIG_CSL_BTD_1**	**RBPJ**	**Q06330‡**
	EBNA2	**Q3KSV2**	**LIG_CSL_BTD_1**	**RBPJ**	**Q06330**
	E1A	**P03254**	**LIG_CtBP_PxDLS_1**	**CTBP1**	**Q13363‡**
	P1234	**P03317**	**LIG_G3BP_FGDF_1**	**G3BP1**	**Q13283‡**
	P1234	**P03317**	**LIG_G3BP_FGDF_1**	**G3BP2**	**Q9UN86‡**
	UL48	**P06492**	**LIG_HCF-1_HBM_1**	**HCFC1**	**P51610**
	VACWR159	P68619	LIG_KLC1_WD_1	KLC1	Q07866‡
	gag	**P03347**	**LIG_LYPXL_L_2**	**PDCD6IP**	**Q8WUM4**
	E1A	**P03255**	**LIG_MYND_1**	**ZMYND11**	**Q15326‡**
	EBNA2	**P12978**	**LIG_MYND_1**	**ZMYND11**	**Q15326**
	E6	P03126	LIG_PDZ_Class_1	TAX1BP3	O14907‡
	E6	P03126	LIG_PDZ_Class_1	DLG1	Q12959‡
	E6	P03126	LIG_PDZ_Class_1	SCRIB	Q14160‡
	E6	**P03126**	**LIG_PDZ_Class_1**	**MAGI1**	**Q96QZ7‡**
	E6	P06463	LIG_PDZ_Class_1	TAX1BP3	O14907
	E6	**P06463**	**LIG_PDZ_Class_1**	**DLG1**	**Q12959**
	E6	P06463	LIG_PDZ_Class_1	SCRIB	Q14160
	E6	P06463	LIG_PDZ_Class_1	MAGI1	Q96QZ7
	E4	P89079	LIG_PDZ_Class_1	DLG1	Q12959
	gag	P03347	LIG_PTAP_UEV_1	TSG101	Q99816
	Segment-10	**P08363**	**LIG_PTAP_UEV_1**	**TSG101**	**Q99816‡**
	gag	P18095	LIG_PTAP_UEV_1	TSG101	Q99816‡
	T-antigen	**P03077**	**LIG_PTB_Phospho_1**	**SHC1**	**P29353‡**
	E1A	**P03255**	**LIG_Rb_pABgroove_1**	**RB1**	**P06400‡**
	polyprotein	P26662	LIG_SH3_2	GRB2	P62993
	UL37	**P10221**	**LIG_TRAF6**	**TRAF6**	**Q9Y4K3**
VirHostNet2.0	LT-Ag	B8ZX42	LIG_Rb_LxCxE_1	RB1	P06400
	LT-Ag	**P03070**	**LIG_Rb_LxCxE_1**	**RB1**	**P06400**
	T-antigen	**P03077**	**LIG_PTB_Phospho_1**	**SHC1**	**P29353‡**
	E6	P03126	LIG_PDZ_Class_1	TAX1BP3	O14907
	E6	P03126	LIG_PDZ_Class_1	MPDZ	O75970
	E6	P03126	LIG_PDZ_Class_1	TJP1	Q07157
	E6	P03126	LIG_PDZ_Class_1	DLG1	Q12959
	E6	P03126	LIG_PDZ_Class_1	SCRIB	Q14160
	E6	P03126	LIG_PDZ_Class_1	DLG2	Q15700
	E6	P03126	LIG_PDZ_Class_1	MAST2	Q6P0Q8

	E6	P03126	LIG_PDZ_Class_1	MAGI1	Q96QZ7
	E7	**P03129**	**LIG_Rb_LxCxE_1**	**RB1**	**P06400‡**
	EBNA6	**P03204**	**LIG_CSL_BTD_1**	**RBPJ**	**Q06330**
	LMP1	P03230	LIG_TRAF2_2	TRAF2	Q12933
	E1A	**P03254**	**LIG_CtBP_PxDLS_1**	**CTBP1**	**Q13363‡**
	E1A	**P03255**	**LIG_Rb_LxCxE_1**	**RB1**	**P06400‡**
	E1A	**P03255**	**LIG_Rb_pABgroove_1**	**RB1**	**P06400‡**
	E1A	**P03255**	**LIG_MYND_1**	**ZMYND11**	**Q15326‡**
	P1234	**P03317**	**LIG_G3BP_FGDF_1**	**G3BP1**	**Q13283‡**
	P1234	**P03317**	**LIG_G3BP_FGDF_1**	**G3BP2**	**Q9UN86‡**
	PB2	P03428	TRG_NLS_Bipartite_1	KPNA1	P52294
	E6	P06463	LIG_PDZ_Class_1	TAX1BP3	O14907
	E6	P06463	LIG_PDZ_Class_1	MPDZ	O75970
	E6	P06463	LIG_PDZ_Class_1	DLG1	Q12959
	E6	P06463	LIG_PDZ_Class_1	SCRIB	Q14160
	E6	P06463	LIG_PDZ_Class_1	DLG2	Q15700
	E6	P06463	LIG_PDZ_Class_1	MAST2	Q6P0Q8
	E6	**P06463**	**LIG_PDZ_Class_1**	**MAGI1**	**Q96QZ7‡**
	Segment-10	**P08363**	**LIG_WW_1**	**NEDD4**	**P46934‡**
	Segment-10	P08363	LIG_WW_1	ITCH	Q96J02
	Segment-10	**P08363**	**LIG_PTAP_UEV_1**	**TSG101**	**Q99816‡**
	gag	P14349	LIG_PTAP_UEV_1	TSG101	Q99816
	gag	P18095	LIG_PTAP_UEV_1	TSG101	Q99816
	UL56	**P28282**	**LIG_WW_1**	**NEDD4**	**P46934‡**
	VACWR159	P68619	LIG_KLC1_WD_1	KLC1	Q07866‡

Known DMIs are shown in bold.

‡ DMIs captured by both datasets.

4.5.4. DMI prediction using known instances and Pfam domains

The ELMc-Domain strategy was used to further increase the number of predicted DMI. For this purpose, known viral instances were linked to Pfam domain containing human proteins via ELMs. PHISTO dataset captured 76 non-redundant DMIs where 18 unique motifs of 27 viral sequences interacted with 17 distinct domains of 51 host proteins **(Figure 4.4, Figure 4.5).** The FDR rate of these prediction was 0.124. VirHostNet2.0 PPI data captured 114 (106 NR) DMIs where 15 unique ELMs of 17 viral sequences interacted with 12 distinct domains of 60 host proteins **(Figure 4.4).** The FDR rate of these predictions was 0.0862 **(Figure 4.5).** PHISTO showed 8.04x enrichment and VirHostNet2.0 showed 11.6x enrichment. Significant enrichment was observed for both datasets (P-value <0.001) **(Figure 4.6C)**.

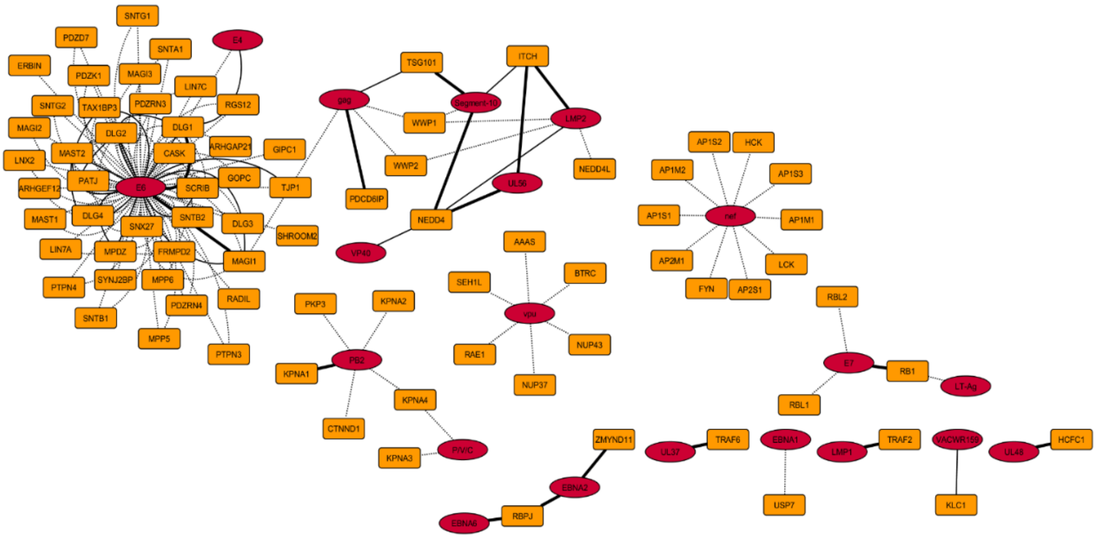


Figure 4.4. Domain-motif interaction (DMI) network of known and predicted DMIs.

DMIs resolved at the PPI level. Red ellipses, viral proteins. Orange rectangles, human proteins**.** Thick solid black lines, DMIs captured using ELMi-Protein strategy. Thin solid black lines, DMIs captured using ELMc-Protein strategy. Black dotted lines, DMIs captured using ELMc-Domain strategy.

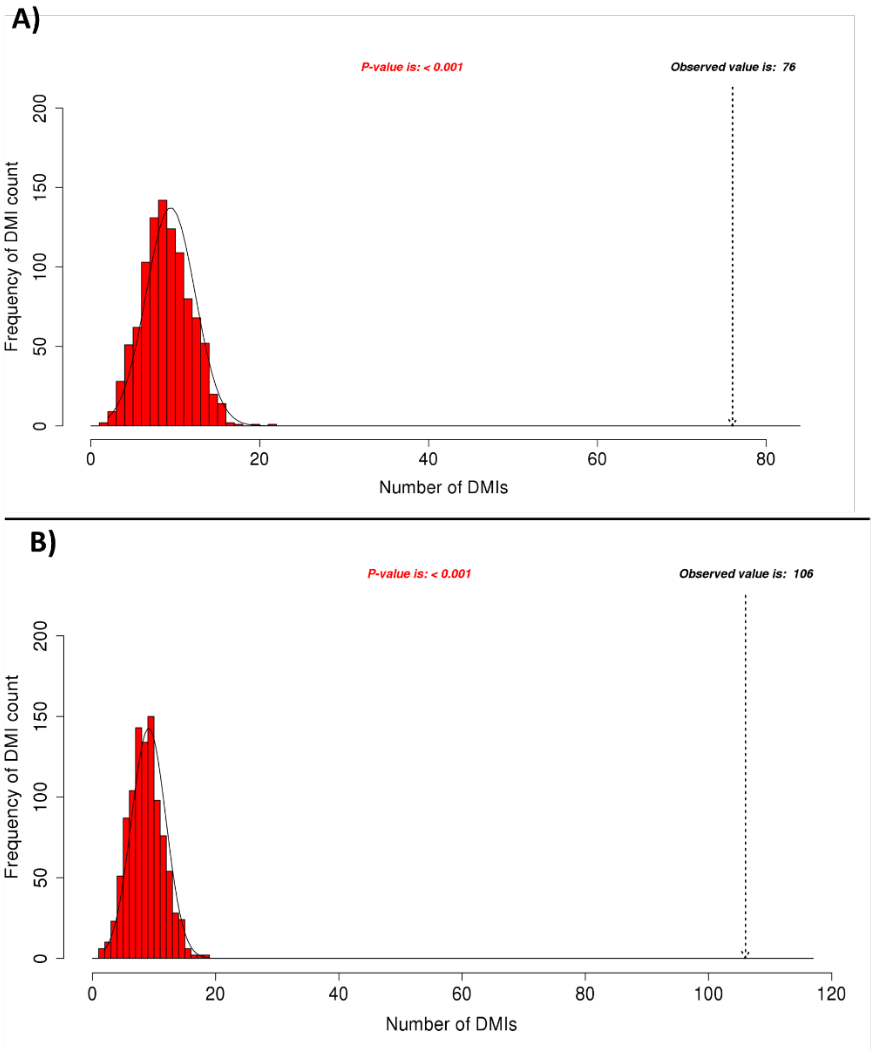


Figure 4.5. SLiMEnrich histograms of observed and expected DMI counts in vhPPI datasets.

A) Absolute number of NR DMI in PHISTO dataset.

B) Absolute number of NR DMI in VirHostNet2.0 dataset. Frequency bars indicate the number of randomised PPI datasets returning a given number of DMIs. The dotted arrow indicates the observed number of DMIs.

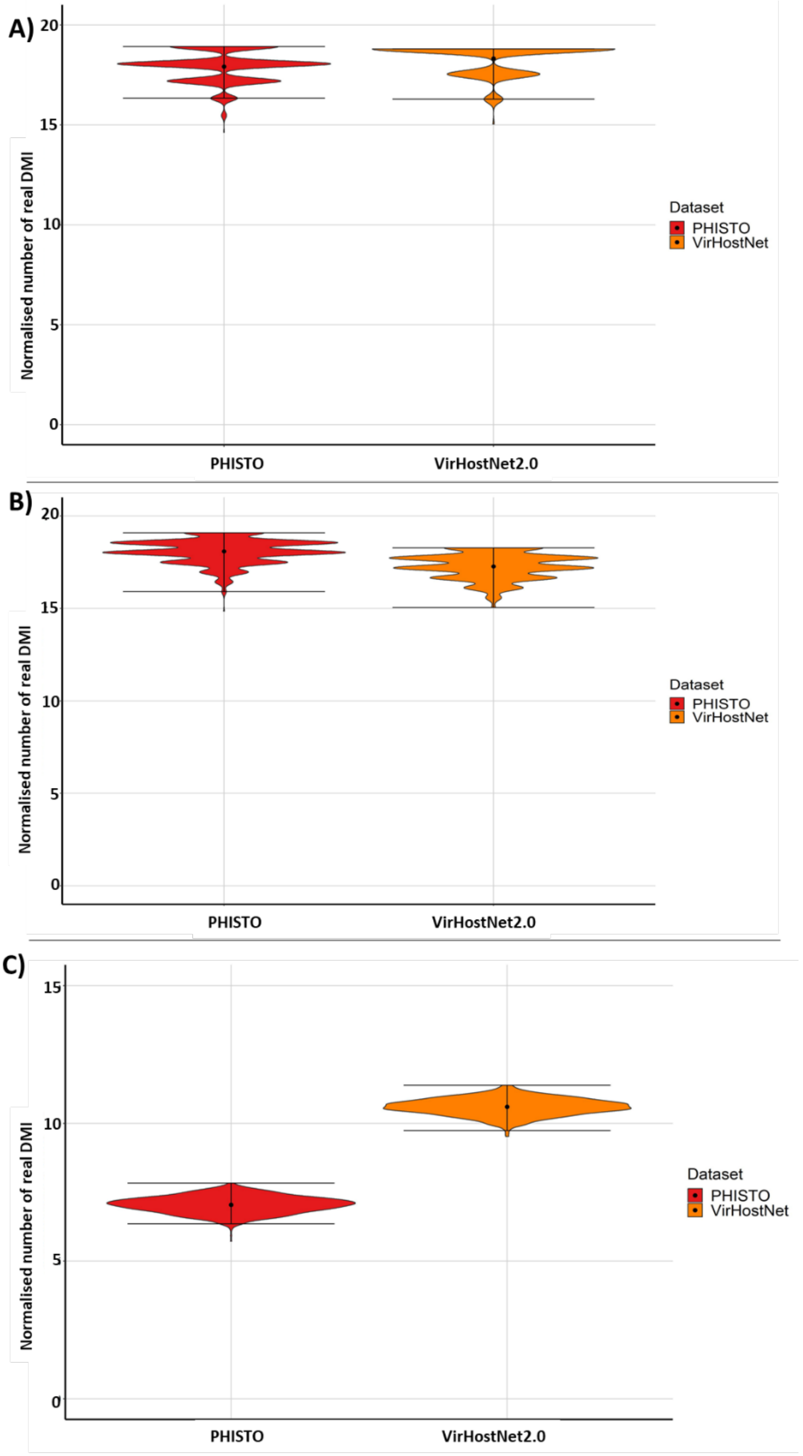


Figure 4.6. Normalised number of real DMIs (DMI $_{Real}$ = DMI $_{Obs}$ – DMI $_{Ran}$).

Real DMIs captured by the vhPPI datasets over 1000x randomisations using ELMc-Domain strategy. Y-axis shows the normalised number of real DMIs. PHISTO enrichment is shown in red and VirHostNet2.0 enrichment is shown in orange. Black dot inside bar shows the median of the real DMIs.

A) Normalised number of real DMIs captured using ELMi-Protein strategy, **B)** Normalised number of real DMIs captured using ELMc-Protein strategy where known viral instances were used, **C)** Normalised number of real DMIs captured using ELMc-Domain strategy where known viral instances were used along with Pfam domains.

On a general note, the predicted DMIs had fewer viral proteins interacting with higher number of (~2x for PHISTO and ~4x for VirHostNet2.0) host proteins **(Figure 4.7).** Most of the predicted DMIs in PHISTO database were in VirHostNet2.0 database **(Figure 4.8A).** Around 55% of viral proteins in DMIs captured by PHISTO were in VirHostNet2.0 database **(Figure 4.8B).** Similarly, ~50% of the human proteins in DMIs captured by PHISTO were in VirHostNet2.0 database **(Figure 4.8C).**

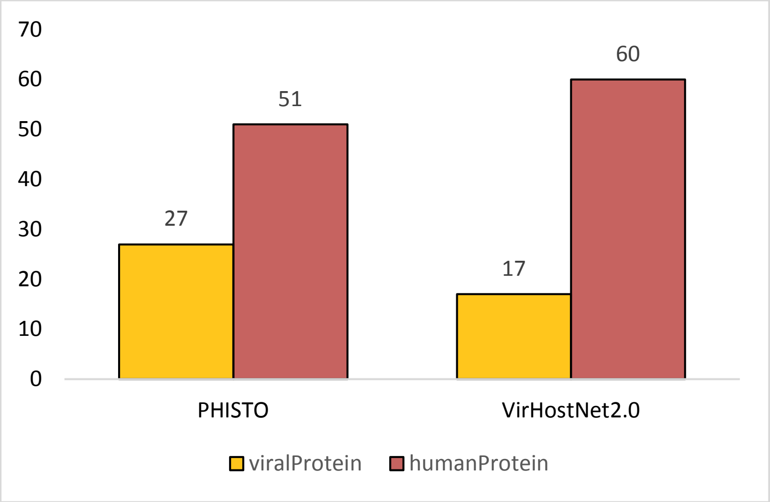


Figure 4.7. Number of viral Proteins and human proteins in DMIs.

Viral and human proteins involved in DMIs predicted using ELMc-Domain strategy where known SLiM instances were used. Yellow represents viral Proteins and red represents human proteins.

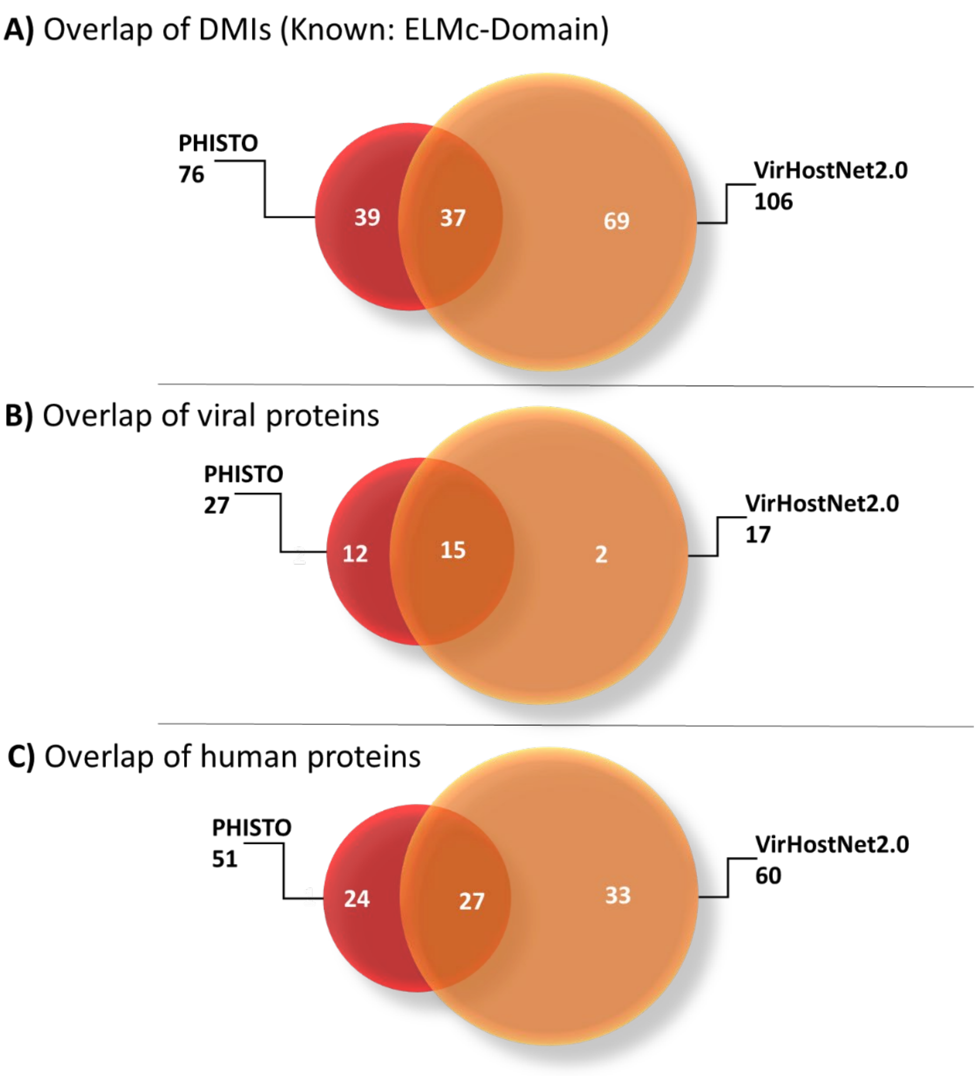


Figure 4.8. Overlap of PHISTO and VirHostNet2.0 datasets.

Red circles represent PHISTO and orange circle represents VirHostNet2.0 datasets. **A)** Overlap of DMIs captured by PHISTO and VirHostNet2.0 datasets. Numbers shows the non-redundant DMIs captured using known SLiMs via ELMc-Domain strategy. **B)** Overlap of non-redundant viral proteins involved in DMIs, **C)** Overlap of non-redundant human proteins involved in DMIs.

4.5.5. DMI prediction using predicted viral instances of known ELMs

To predict new candidates for molecular mimicry, SLiM instances of known ELMs were predicted in all viral proteins using SLiMProb v2.5.1 (Edwards and Palopoli 2015) with the disordered masking feature (IUPred score >= 0.2) (Hagai, Azia et al. 2011). The predicted SLiMs were then used to predict DMIs using ELMc-Protein strategy. Both PHISTO and VirHostNet2.0 datasets still showed significant enrichment for DMIs **(Figure 4.9)**. PHISTO predicted 294 non-redundant DMIs where 46 unique motifs of 202 viral sequences interacted with 54 host proteins. The FDR of these predictions was 0.4154 **(Figure 4.9A).** VirHostNet2.0 returned 196 (183 NR) DMIs where 42 unique ELMs of 113 viral sequences interacted with 53 host proteins. The FDR of these predictions was 0.3923 **(Figure 4.9B).** PHISTO showed 2.40x enrichment and VirHostNet2.0 showed 2.54x enrichment, both highly significantly enriched versus random data (P-value <0.001) **(Figure 4.11A).**

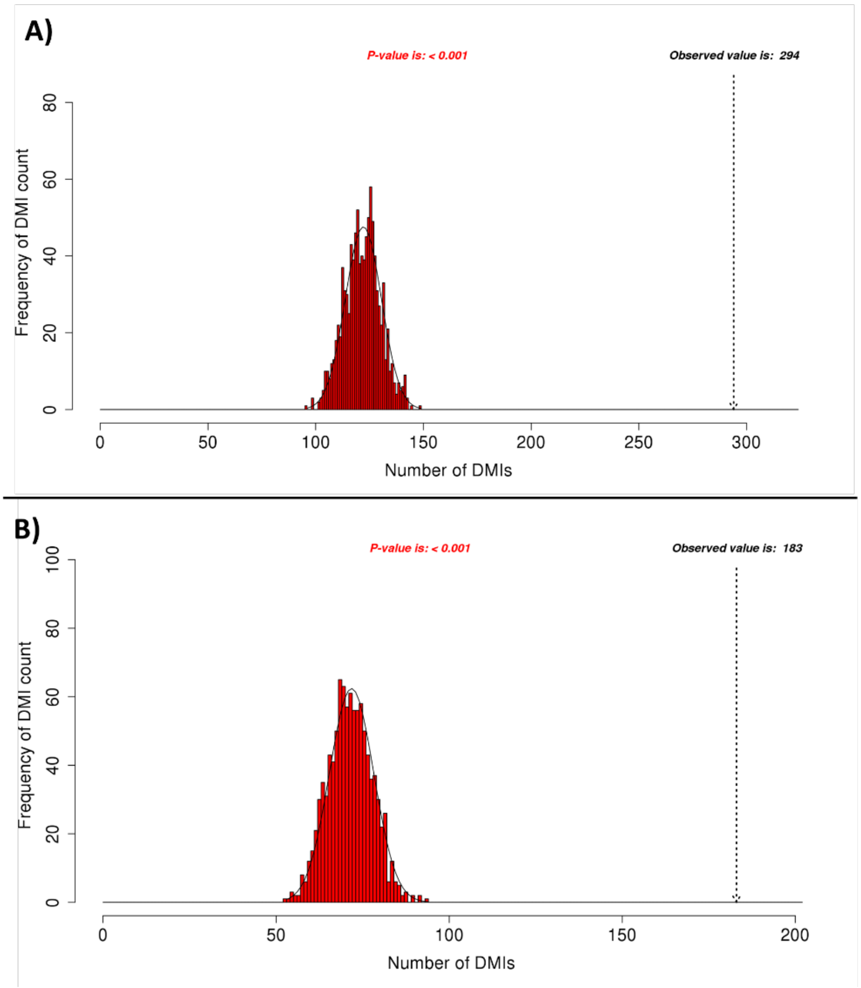


Figure 4.9. SLiMEnrich histograms of observed and expected DMI counts in vhPPI datasets.

Expected DMI counts using ELMc-Protein strategy where predicted SLiMs were used, **A)** Absolute number of DMI count in PHISTO dataset, **B)** Absolute number of DMI count in VirHostNet2.0 dataset. Frequency bars indicate the number of randomised PPI datasets returning a given number of predicted DMIs. The dotted arrow indicates the observed number of DMIs.

4.5.6. DMI prediction using predicted viral SLiMs and Pfam domains

Finally, predicted viral SLiMs were linked to human proteins via ELM-binding Pfam domains. This introduction of noise in DMI network drastically increased DMI number while lowering the overall enrichment. PHISTO returned 6,498 (2,728 NR) DMIs, where 120 unique motifs of 582 viral sequences interacted with 56 distinct domains of 541 host proteins. The FDR of these predictions was 0.7507**(Figure 4.10A).** VirHostNet2.0 PPI returned 3,025 (1,131 NR) DMIs where 111 unique ELMs of 274 viral sequences interacted with 51 distinct domains of 457 host proteins **(Figure 4.10B)**. The FDR of these predictions was 0.7843**.** PHISTO showed 1.33x enrichment and VirHostNet2.0 showed 1.27x enrichment **(Figure 4.11B).** Both PHISTO and VirHostNet2.0 were still significantly capturing DMIs (P-value <0.001).

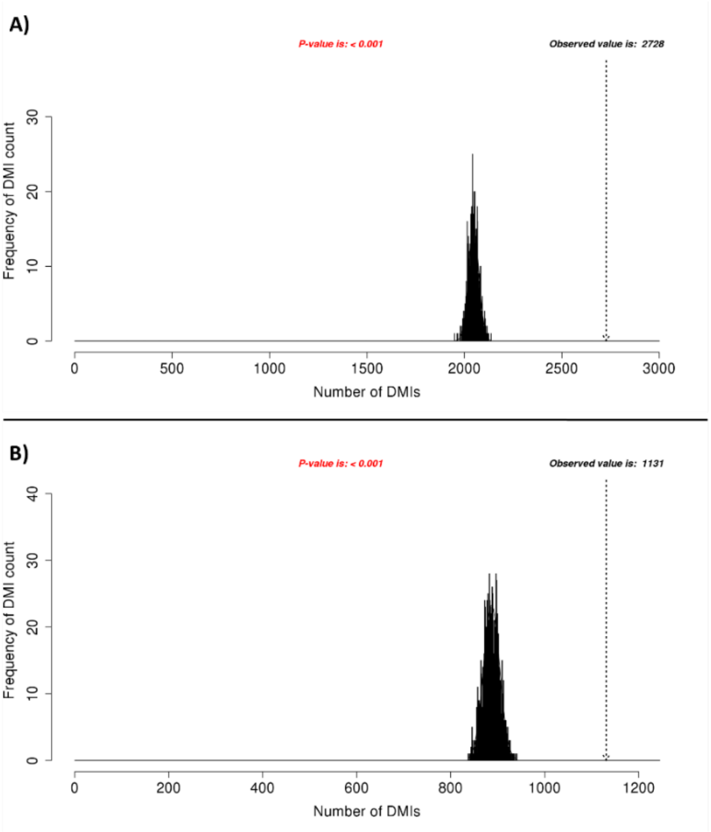


Figure 4.10. SLiMEnrich histograms of observed and expected DMI counts in vhPPI datasets.

Expected DMIs using ELMc-Domain strategy where predicted SLiMs were used, **A)** Absolute number of DMI count in PHISTO dataset, **B)** Absolute number of DMI count in VirHostNet2.0 dataset. Frequency bars indicate the number of randomised PPI datasets returning a given number of predicted DMIs. The dotted arrow indicates the observed number of DMIs.

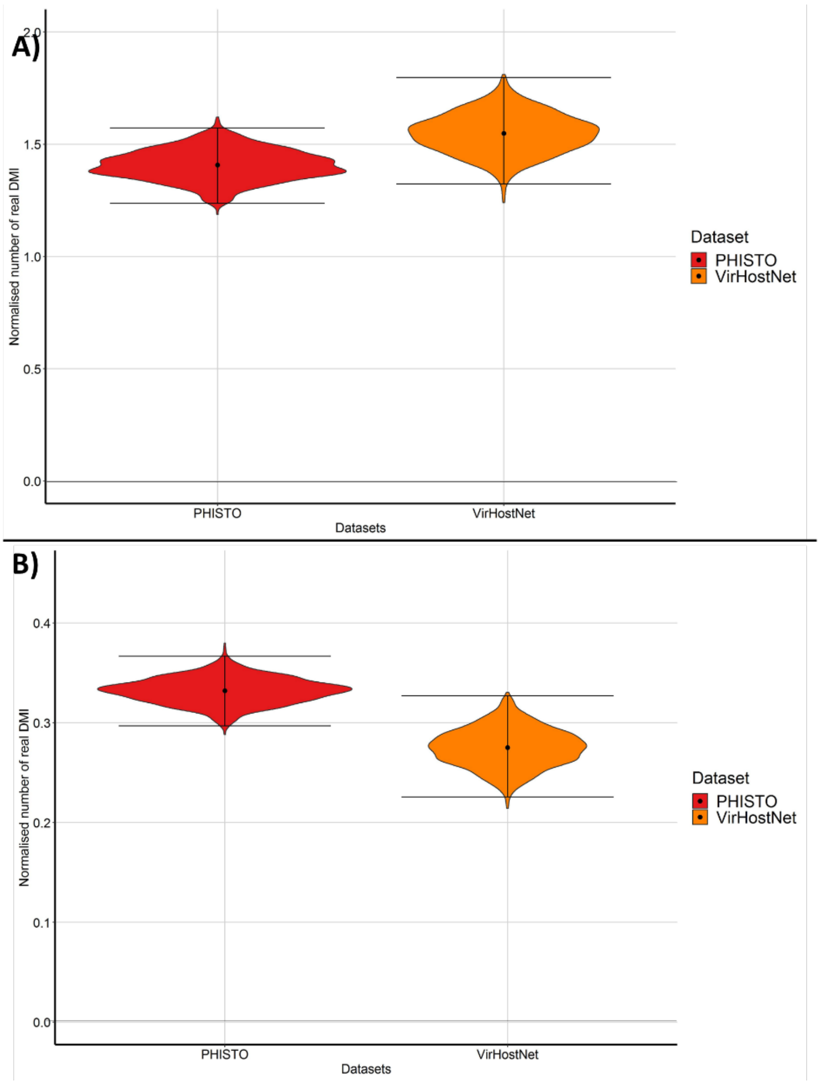


Figure 4.11. Normalised number of real DMIs captured by the vhPPI datasets over 1000x randomisations.

PHISTO enrichment is shown in red and VirHostNet2.0 enrichment is shown in orange. Black dot inside bar shows the median of the real DMIs. Y-axis shows the normalised number of real DMIs. **A)** Normalised number of real DMIs captured using ELMc-Protein strategy where new instances of known ELMs were used, **B)** Normalised number of real DMIs captured using ELMc-Domain strategy where new instances of known ELMs were used.

4.5.7. Molecular mimicry in different classes of virus

To date, there has been no study to analyse different viral subtypes based on their genetic material to see how they perturb host cellular machinery for their regulatory functions and infection cycle. Therefore, the main objective was to see how different viruses tend to interact with host proteins through SLiMs based on their genetic material.

4.5.7.1. DMI Prediction in different viral subtypes

PHISTO and VirHostNet2.0 databases were split into different groups based on viral genetic material: RNA single stranded (ssRNA), RNA double stranded (dsRNA), DNA single stranded (ssDNA) and DNA double stranded (dsDNA). SLiMProb v2.5.1 was run on viral proteins of each viral group to predict new SLiM instances of known ELMs using disorder masking. Predicted SLiM instances were then used to predict new DMIs using the ELMc-Domain strategy. The reason of choosing this strategy was to see how predicted viral motifs hijack host cellular machinery through mimicking specific domain interaction partners. The MOD and CLV ELM classes, which are involved in post-translational modifications, tend to be low complexity and/or have a high prevalence of interacting domains in the human proteome, which makes them generally poorly enriched for DMI **(Figure 3.13)**. These classes were excluded from the analysis to reduce the false discovery rate and focus on DMIs that are more likely to be true positive by reducing noise in the network.

1. **DMI Prediction in RNA viruses**

Both databases had more ssRNA interactions than dsRNA interactions **(Table 4.3).** PHISTO database had 58 dsRNA vhPPIs, of which only 1 was predicted to be a DMI with 2.3x enrichment **(Figure 4.12A, Table 4.3)**. VirHostNet2.0 had 165 dsRNA vhPPIs, of which only 4 were predicted as DMIs with 9.78x enrichment **(Figure 4.12A, Table 4.3).** The ssRNA vhPPIs of both databases predicted more DMIs, but with lower enrichment in each case. Out of 10,389 vhPPIs in PHISTO, 448 were predicted DMIs with 1.27x enrichment and out of

14,892 vhPPIs in VirHostNet2.0, 208 were predicted DMIs with 3.37x enrichment **(Figure 4.12, Table 4.3).**

Table 4.3. DMI prediction and enrichment analysis of viral subtypes.

Dataset	Viral subtype	vhPPIs[1]	potDMI[2]	Predicted DMI[3]	E-score	FDR
PHISTO	**dsRNA**	58	2	1*	2.30	0.434
	ssRNA	10,389	58,209	448**	1.27	0.783
VirHostNet2.0	**dsRNA**	165	28	4**	9.78	0.102
	ssRNA	14,892	14,131	208**	3.37	0.296
PHISTO	**dsDNA**	212	335	17*	1.30	0.769
	ssDNA	62	60	8*	0.92	1.000
VirHostNet2.0	**dsDNA**	29,532	166,264	561**	2.26	0.441
	ssDNA	640	295	16**	1.90	0.526

*P-value < 0.01

**P-value < 0.001

1. Non-redundant vhPPIs.
2. Non-redundant number of all possible DMIs between a motif containing protein and Pfam domain containing protein available in vhPPIs.
3. Non-redundant predicted DMIs excluding PTMs.

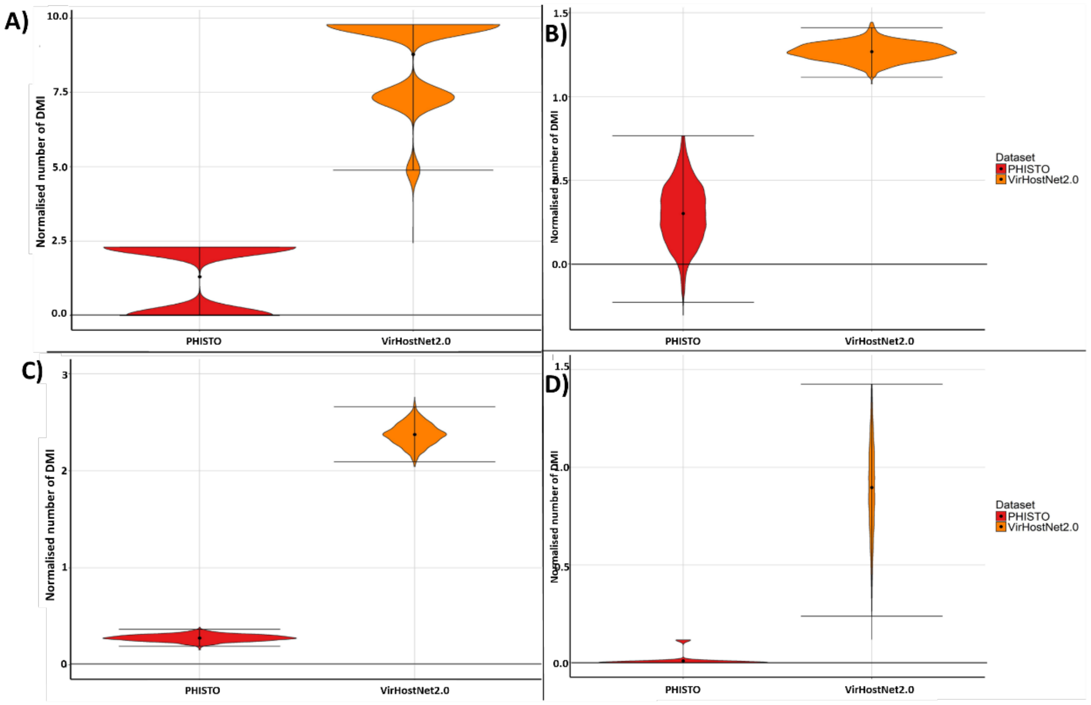


Figure 4.12. Normalised number of real DMIs predicted using viral subtype PPIs available in PHISTO and VirHostNet2.0 using ELMc-Domain strategy.

A) Normalised number of real DMIs predicted over 1000x randomisations using predicted SLiM instances of dsRNA viruses, **B)** Normalised number of real DMIs predicted over 1000x randomisations using predicted SLiMs of dsDNA viruses. **C)** Normalised number of real DMIs predicted over 1000x randomisations using predicted SLiM instances of ssRNA viruses, **B)** Normalised number of real DMIs predicted over 1000x randomisations using predicted SLiMs of ssDNA viruses.

Gene ontology enrichment analysis was performed to see which biological processes were being disrupted by viral proteins upon targeting human proteins. The human proteins targeted by ssRNA viral proteins were enriched in regulation and metabolic related processes (p-value < 0.001). The targeted proteins were also involved in binding (i.e. protein binding and calcium binding) functions and activities like transferase and kinase. The targeted proteins were mostly enriched in cytoplasm (more specifically in cytosol) and cytoskeleton **(Figure 4.13).**

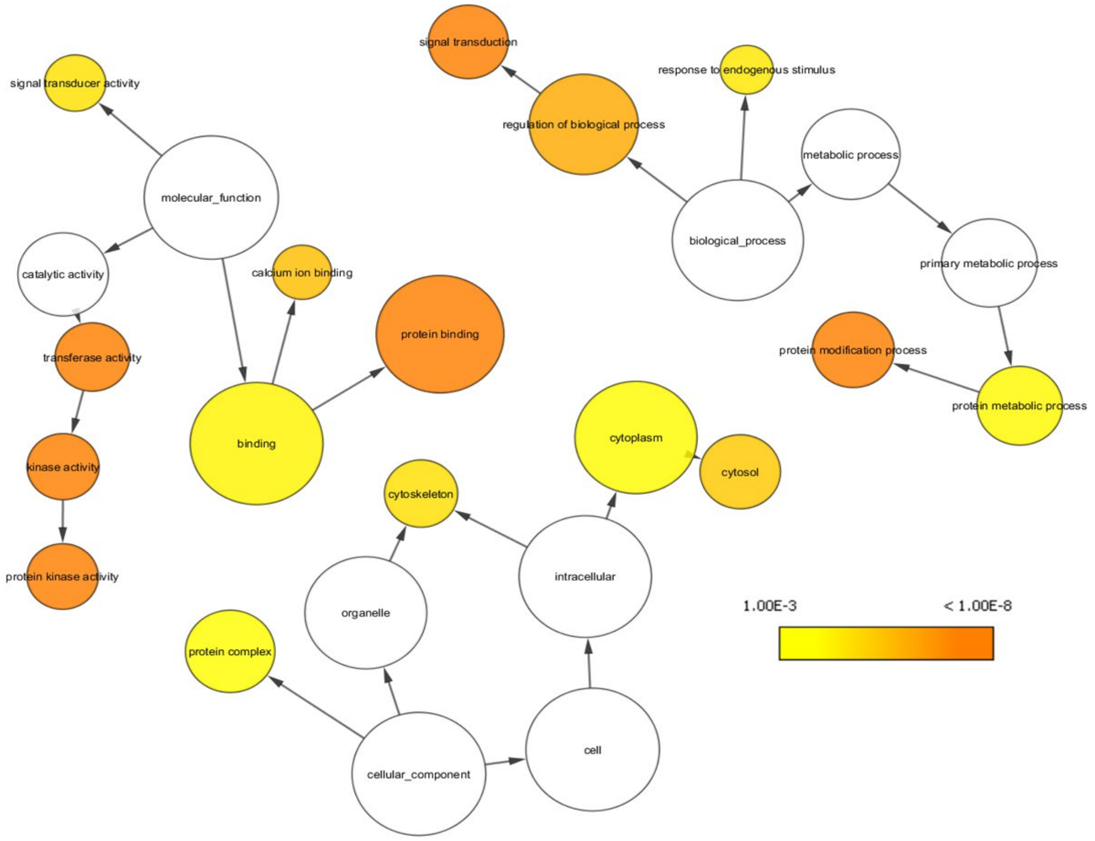


Figure 4.13. Gene ontology enrichment analysis of human proteins targeted by ssRNA viral proteins.

Size of the circle shows number of proteins; darker shaded circles represent more enrichment while lighter shades represents lower enrichment.

On the other hand, human proteins targeted by dsRNA viral proteins were involved in symbiosis, reproduction, regulatory, metabolic, modification processes and catalytic activities. These proteins were mostly enriched in plasma membrane **(Figure 4.14).**

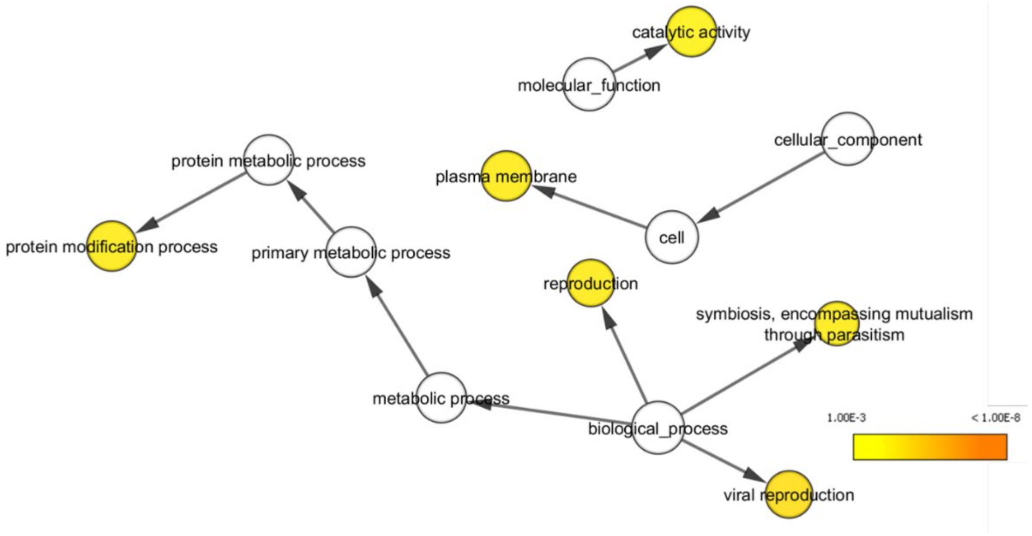


Figure 4.14. Gene ontology enrichment analysis of human proteins targeted by dsRNA viral proteins.

Size of the circle shows number of proteins; darker shaded circles represent more enrichment while lighter shades represents lower enrichment.

2. DMI Prediction in DNA viruses

PHISTO database had only 212 dsDNA interactions, of which only 17 were being mediated by SLiMs **(Table 4.3)**. VirHostNet2.0 had large number of dsDNA interactions (29,532 vhPPI) of which only a small fraction (561) were DMI. Both databases had lower number of ssDNA interactions in comparison to dsDNA. PHISTO had 62 vhPPIs, of which 8 were DMIs. VirHostNet2.0 had 640 vhPPIs, of which 16 were DMIs**.** dsDNA vhPPIs were more enriched than ssDNA vhPPIs in terms of capturing DMIs **(Figure 4.12**, **Table 4.3).**

Gene ontology enrichment analysis showed that most of the human proteins targeted by dsDNA viral proteins were primarily enriched in metabolic, regulation, signal transduction and cell cycle processes. These proteins were mainly enriched in nucleoplasm and cytosol. These human proteins were mostly involved in molecular functions of protein binding, nucleotide binding and activities like kinase and transferase **(Figure 4.15).**

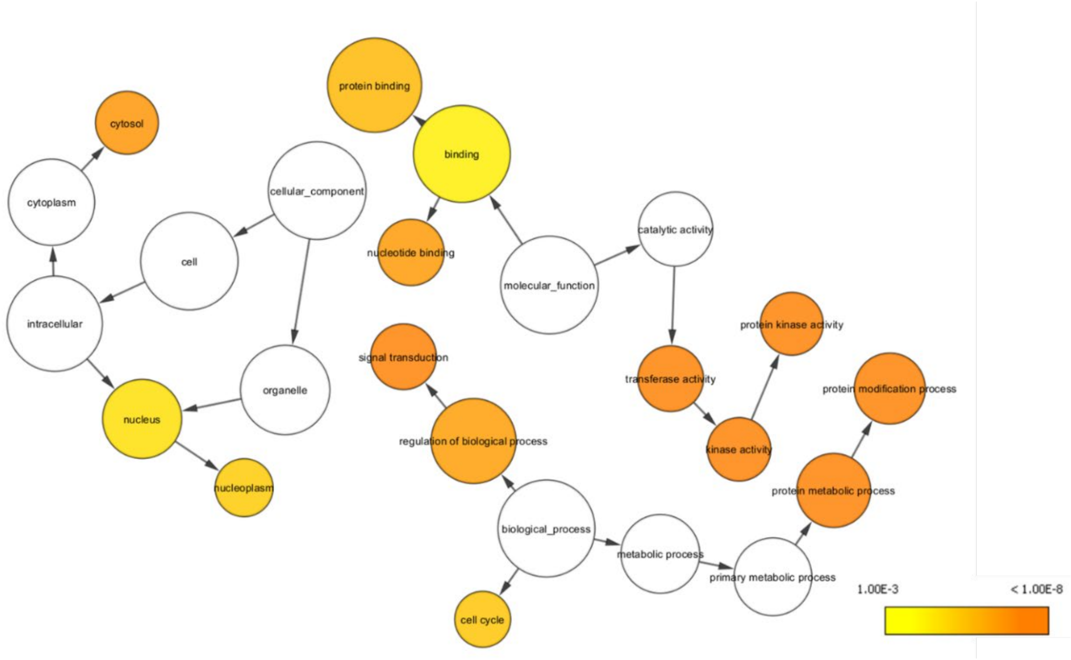


Figure 4.15. **Gene ontology enrichment analysis of human proteins targeted by dsDNA viral proteins.**

Size of the circle shows number of proteins; darker shaded circles represent more enrichment while lighter shades represents lower enrichment.

On the other hand, human proteins targeted by ssDNA viral proteins were found to be enriched in metabolic related processes and kinase activity **(Figure 4.16).**

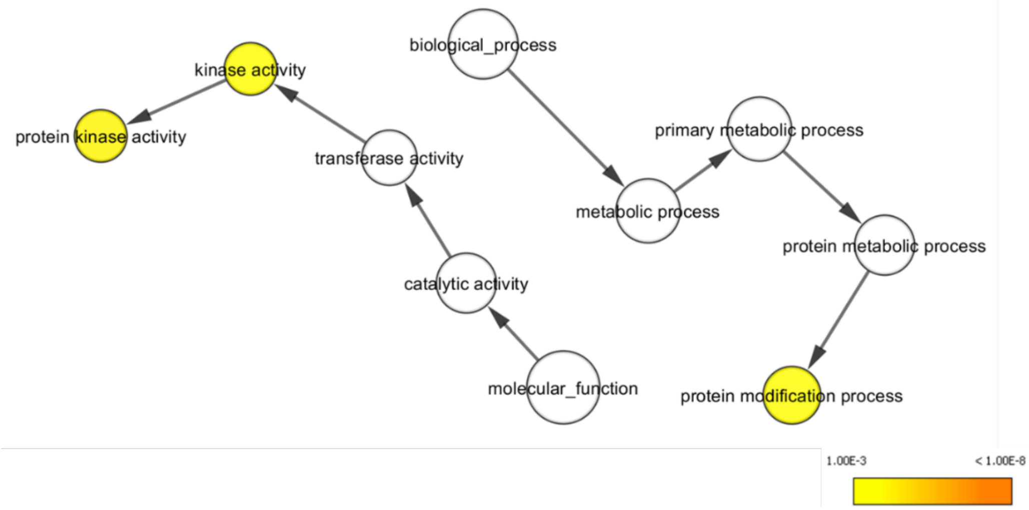


Figure 4.16. Gene ontology enrichment analysis of human proteins targeted by ssDNA viral proteins.

Size of the circle shows number of proteins; darker shaded circles represent more enrichment while lighter shades represents lower enrichment.

4.5.8. *De-novo* SLiM discovery

Previous analysis revealed that both human and viral-human interactomes were capturing DMIs with significant enrichment and so these data were used for *de-novo* SLiM discovery. A total of 12,139 datasets were generated for vhPPI pairs; for each pair the dataset contained a single viral protein and all the human interactors of the viral protein's human interaction partner. The generated datasets were fed to QSLiMFinder where viral proteins were treated as a query to predict SLiMs. As per the SLiM discovery criteria (Palopoli, Lythgow et al. 2015), datasets which had too few or too many UPCs (Unrelated protein clusters) were disregarded from the analysis. A total of 1,857 significant datasets (p-value <0.1) returned 2,564 motifs in the real group. Given the large number of datasets, and previous observation that QSLiMFinder is not a stringent as SLiMFinder (Palopoli, Lythgow et al. 2015), it was decided to focus on results with the more stringent significance thresholds of ≤0.01. Motifs with overlapping patterns and instances are clustered into "clouds" in QSLiMFinder. A total of 300 datasets returned motifs (308 clouds, 177 motif patterns) at P ≤0.01 **(Figure 4.17A).**

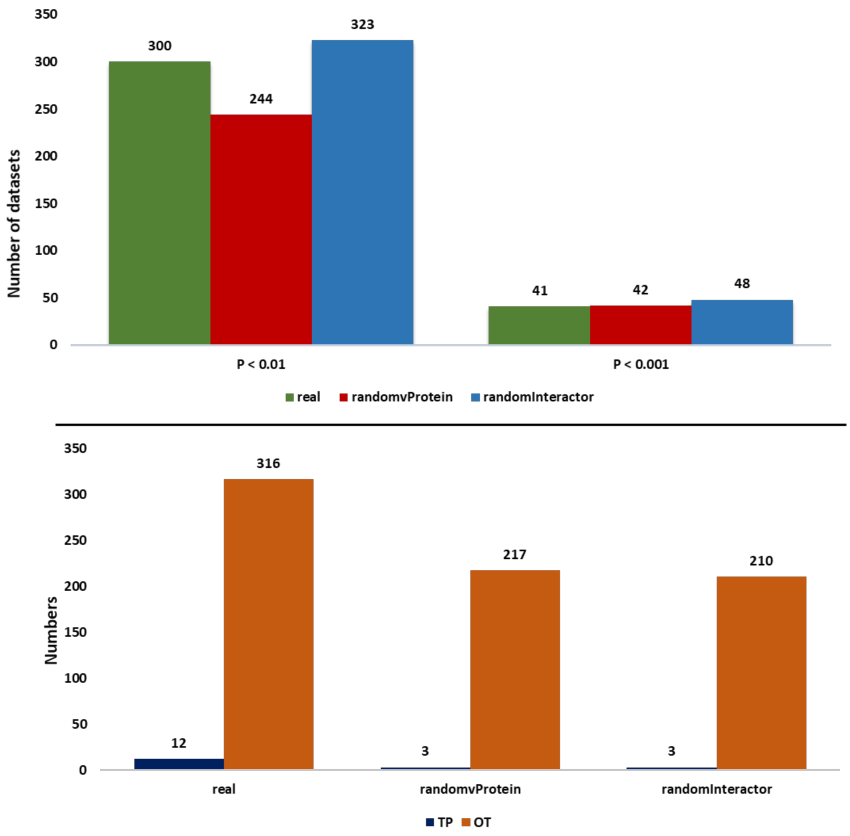


Figure 4.17. QSLiMFinder and CompariMotif analysis of *de-novo* SLiM mimicry prediction.

- **A)** Number of datasets returning SLiMs vs significant p-value calculated by SLiMChance. X-axis shows the P-value cut-off and y-axis shows the number of datasets returning SLiMs with cut-off P-values. Real datasets are shown in green, control group 1 (randomised viral proteins) is shown in red and control group 2 (randomised human interactors) is shown in blue.
- **B)** Number of true positives (TP clouds), hub proteins having annotations in ELM, number of off-targets (OTs) in real and control groups.

To further assure that the pipeline was effective in terms of *de-novo* discovery of viral SLiM mimicry, two random control groups were simulated where first group had randomised viral proteins, and second group had randomised human interactor proteins. A total of 1,683 significant randomvProtein datasets returned 2,416 individual motifs; 244 datasets returned enriched sequence patterns (262 clouds, 217 motif patterns) at P ≤0.01. 1,813 significant randomInteractor datasets returned 2,364 individual motifs; 323 datasets returned enriched sequence patterns (366 clouds, 158 motif patterns) at P ≤0.01. Only a small proportion of significant datasets (~0.4%) returned motifs at P ≤0.001. **(Figure 4.17A).**

Next, discovered SLiMs were compared with SLiMs available in ELM to see how many of the predicted SLiMs were overlapping with the known SLiMs. This was done using CompariMotif v3.3.1 which compares two lists of regular expression motifs with each other to find overlap and relationships between them. A motif was regarded as a TP if the hub protein was known to interact with (or contains a domain that interacts with) the identified ELM. 12 motifs in real data, 3 in randomvProtein group and 3 in randomInteractor group were found to be true positives. Out of 2,564 predicted SLiMs in real data (p-value cutoff ≤0.1), 316 were regarded as OTs as they matched pattern with motifs in ELM based on match criteria (i.e. MatchIC ≥ 2.5 or normalised IC ≥ 1.0), 217 in control group 1 matched pattern with ELM , and 210 enriched sequence patterns in control group 2 matched pattern with motifs in ELM **(Figure 4.17B)**.

4.6. Discussion

Virus interacts with their hosts through establishing protein-protein interactions (vhPPIs). (Garamszegi, Franzosa et al. 2013). Most of the time viruses use vhPPIs to mimic host proteins: a viral protein having sequence or structural similarity as that of host protein binds with the host protein binding partner and disrupts host cellular pathways (Davey, Trave et al. 2011; Berlow, Dyson et al. 2018). Viral mimicry is often achieved through short linear motifs (SLiMs) which mimic host protein SLiMs and establishes low affinity domain-motif interactions (DMIs) with binding proteins (Benedict, Norris et al. 2002; Van Roey, Uyar et al. 2014). For example, E6 protein of human papilloma virus (HPV) interacts with PDZ domain containing proteins (Ganti, Broniarczyk et al. 2015). The number of host proteins is quite large in many organisms, which makes it expensive to determine vhPPIs through *in-vitro/in-vivo* experiments. Moreover, the transient nature of SLiM mediated interactions further complicates their detection through experimental techniques (Becerra, Bucheli et al. 2017). Computational methods of predicting viral mimicry could be an inexpensive and ideal way to predict interactions.

In Chapter 2, a new pipeline known as SLiMEnrich was developed, which not only assesses whether a PPI data is good for capturing SLiM mediated interactions but also predicts new interactions. In this chapter, I have applied SLiMEnrich to first assess whether vhPPI data is enriched for DMIs and which high-throughput (two hybrid and affinity purification) method is better for predicting DMIs in vhPPIs. Once assured that vhPPI data was indeed capturing DMIs, my next aim was to see how different viral subtypes perturb host cellular machinery through DMIs. Lastly, I combined the full vhPPI data with human-human PPI data for *de-novo* prediction of human SLiM mimicry in viruses

4.6.1. vhPPIs capture Domain-Motif Interactions (DMIs)

Chapter 3 revealed that human interactomes, including those captured by high throughput PPI detection methods (Y2H and AP-MS), are significantly enriched for DMIs versus the random expectation **(Table 3.1).** As such, these PPI data are a legitimate source for discovering new DMIs. In this chapter, the focus is virus-host PPI (vhPPI), raising the question whether this observation was also true for vhPPI data. Two large-scale sources of vhPPI data are PHISTO (Durmus Tekir, Cakir et al. 2013) and VirHostNet2.0 (Guirimand, Delmotte et al. 2015). Both capture DMIs with significant enrichment (Table 4.2, Figure 4.3). Both PHISTO and VirHostNet2.0 datasets captured small number known DMIs: 22 known DMIs were captured by PHISTO and 16 by VirHostNet2.0. On a face value, this looks disappointing, but most of the 1,442 DMIs annotated in ELM (Gouw, Michael et al. 2017) are reported in human interactome. So far only few DMIs (i.e. 85 vhDMIs) have been reported in viruses, which highlights that many DMIs are yet to be discovered in vhPPIs. For known SLiM classes in ELM, the SLiMEnrich approach applied here can potentially predict new candidates of mimicry where viral proteins exploits host functions (section 4.6.3). In general, only ¼ portion of the known vhDMI was being rediscovered using this approach (i.e. 26% in case of PHISTO and 19% in case of VirHostNet2.0) database. The low portion of known vhDMIs returned in vhPPI datasets suggests that atleast 3x as many new DMI are out there which are not in these vhPPI datasets. The low overall return of known viral DMIs also suggests that there are improvements to be made in the compilation and mapping of vhPPI in PHISTO and VirHostNet. As seen in the results, not many known interactions were captured by the vhPPIs, this could be due to same viral families being screened in high-throughput experiments.

4.6.2. High-throughput methods capture virus-host DMIs

Once assured that viral interactomes were enriched in terms of capturing DMIs, I investigated which high-throughput method between Y2H and AP-MS was better at predicting DMIs. The recent progress in PPI detection techniques have led to detection of large scale vhPPI data. Yeast two-hybrid (Y2H) and affinity purification coupled mass spectrometry (AP-MS) are two widely used technologies in terms of detecting vhPPIs (de Chassey, Meyniel-Schicklin et al. 2014). Y2H has been used in 15 high-throughput screens to identify genome wide viral interactomes. The first genome wide vhPPI screens using Y2H technology were done for HCV (de Chassey, Navratil et al. 2008) and Epstein Barr virus (Calderwood, Venkatesan et al. 2007). Moreover, Y2H has also been used to identify vhPPIs focused on specific proteins for example in one study ~12,000 human proteins and 10 influenza virus proteins were used to identify vhPPIs in Influenza virus (Shapira, Gat-Viks et al. 2009). A variation of AP-MS is known as tandem affinity purification (TAP) which has been widely used to identify large numbers of vhPPIs (e.g. (Pichlmair, Kandasamy et al. 2012; Rozenblatt-Rosen, Deo et al. 2012). This technique is being considered good because of its low contamination background and lower rate of false positive interactions (Rigaut, Shevchenko et al. 1999).

The main objective of this analysis was to see whether these methods were good in terms of capturing DMIs from vhPPIs. In **Chapter 3,** both Y2H and AP-MS showed significant DMI enrichment and there was no clear winner between the two methods. They both were capturing significant number of known DMIs **(Table 3.3, Figure 3.2A).** These results were recapitulated for the vhPPI data. High-throughput interactions available in both PHISTO and VirHostNet2.0 vhPPI databases are indeed capturing DMIs **(Table 4.2, Figure 4.3)**. Both AP-MS and Y2H vhPPI screens showed significant enrichment in terms of capturing DMIs. In general, the enrichment trend was similar to the human interactome without a clear "winner" **(Table 3.3, Figure 3.2A)**. The proportion of known vhDMIs captured by vhPPIs was quite low (19-26% known vhDMIs) therefore, to make increase the network, medium stringency filtering was applied where I used ELMc-Protein strategy to see the robustness of the results. As ELMc-Protein strategy adds more DMIs in the results, the enrichment could be more reliable. Both high-throughput methods still showed significant enrichment **(Table 4.1)**. Thus, it can be said that both Y2H and AP-MS screens are capable of capturing virus-host DMIs, and it is appropriate to use both types of data for DMI and SLiM prediction.

4.6.3. DMI prediction using different types of DMI data

Once assured that viral interactomes were capturing DMIs, analysis was extended to predict new DMIs where viruses are likely to mimic host proteins. First, I employed medium stringency strategy of SLiMEnrich (i.e. ELMc-Protein) **(Figure 2.2.)** which led to prediction of 36 DMIs in PHISTO and 34 in VirHostNet2.0 dataset. The FDR associated with these predictions was quite low (i.e. 0.0523 in case of PHISTO and 0.0547 in case of VirHostNet2.0) suggesting that the predicted DMIs might be real and in general, this strategy can be quite useful in identifying real DMIs. All the DMIs belonged to ligand (LIG) and targeting (TRG) ELM type. In total, both databases returned 47 vhDMIs of which 23 were known in ELM and 24 were predicted DMIs **(Table 4.2).** The FDR of ~5% for these predictions makes it likely that the additional 24 DMIs are also real.

A total of 4 DMIs were mediated by LIG_Rb_LxCxE_1 **(Table 4.2)** which is known for interactions with retinoblastoma protein family in human as well as multiple viruses (Liu and Marmorstein 2007; Davey, Trave et al. 2011; Berlow, Dyson et al. 2018). Among these DMIs, 3 were known in ELM. All the viral proteins in these DMIs (i.e. E7 protein of Human papillomavirus type 16, LT-Ag protein of Simian virus 40 (SV40) and E1 protein of Human adenovirus 5) were interacting with RB1 protein in human and were known for mimicry (Chemes, Sanchez et al. 2011). This analysis returned one new DMI between LT-Ag protein of Merkel cell polyomavirus and RB1 protein of human interacting via LxCxE motif which was present in accessible area of LT-Ag protein. As the motif instance is not new therefore, it cannot be ruled out that the predicted DMI may have already been reported in the literature but is not documented in ELM.

A total of 7 DMIs were mediated by LIG_WW_1 **(Table 4.2)** which is a WW domain binding motif (Traweger, Fang et al. 2002) known in different species, including human and viruses (i.e. Human herpesvirus and Ebola virus) (Dinkel, Van Roey et al. 2016). Among these DMIs, 3 were known in ELM and 4 were not annotated in ELM (predicted DMIs). LMP2 protein of Epstein-Barr virus and UL56 protein of Human herpesvirus 2 were interacting ITCH protein in human while Segment-10 protein in Bluetongue virus 10, VP-40 protein in Zaire ebolavirus and UL56 protein in Human herpesvirus 2 were interacting with NEDD4 protein in human. This analysis predicted 2 new DMIs where LMP2 protein in Epstein-Barr virus was interacting with NEDD4 protein in human and Segment-10 protein in Bluetongue virus 10 was interacting with ITCH protein in human via LIG_WW_1 motif which was present in the accessible area of the viral proteins. This reciprocal switch of interaction partners adds confidence that both interactions are genuine, as all proteins involved are known to be involved in mimicry interactions.

A total of 4 DMIs were mediated by LIG_PTAP_UEV_1 **(Table 4.2)** which binds Tsg101 through the N-terminal UEV domain (Schlundt, Sticht et al. 2009). Among these DMIs, 1 was known and 3 were predicted ones. The known DMI was between Segment-10 protein of Bluetongue virus 10 and TSG101 protein of human. The 3 predicted DMIs had interactions among gag proteins of Human spumaretrovirus, Human immunodeficiency virus type 1 and Human immunodeficiency virus type 2 with TSG101 protein of human via LIG_PTAP_UEV_1 motif which was present in the accessible area of the viral proteins.

A total of 16 DMIs were mediated by DMIs were mediated by LIG_PDZ_Class_1 **(Table 4.2)** which is a c-terminal peptide that bind with the PDZ domains. This motif interacts with targeted proteins through beta-augmentation to a beta sheet of PDZ domain in the targeted protein (Hung and Sheng 2002). Among these DMIs, 2 were known DMIs and 14 were predicted ones. One known DMI had interactions between E6 protein of Human papillomavirus type 16 and MAGI1 protein in human. The other known DMI had interaction between E6 protein of Human papillomavirus type 18 and DLG1 protein of human. Out of 13 predicted DMIs, 7 had E6 protein of Human papillomavirus type 16 interacting with 7 distinct human proteins (i.e. TAX1BP3, MPDZ, TJP1, DLG1,SCRIB, DLG2,MAST2). One DMI had E4 protein of Human adenovirus D serotype 9 interacting with DLG1 protein in human while the remaining 6 DMIs had E6 protein of Human papillomavirus type 18 interacting with 6 distinct human proteins (i.e. TAX1BP3, MPDZ, SCRIB, DLG2, MAST2, MAGI1). The motif in interacting viral protein was in the accessible area of the protein.

Only 1 DMI was mediated by LIG_KLC1_WD_1 which is found in cargo proteins and mediates kinesin-1-dependent microtubule transport when bound to the KLC TPR region (Konecna, Frischknecht et al. 2006). The interaction was not known in ELM and had interaction between VACWR159 protein of Vaccinia virus and KLC1 protein of human. The LIG_KLC1_WD_1 motif was in the accessible area of the protein.

A number of known DMIs were returned for different ELMs. Two known DMIs were returned for LIG_CSL_BTD_1. These interactions were between the EBNA2 and EBNA6 proteins of Epstein-Barr virus with the RBPJ proteins in human. One known interaction was returned for LIG_PTB_Phospho_1 where T-antigen protein of Murine polyomavirus had interaction with SHC1 protein in human. One known interaction was mediated by LIG_CtBP_PxDLS_1 where E1A protein of Human adenovirus C serotype 2 was interacting with CTBP1 protein in human. Two known interactions were mediated by LIG_G3BP_FGDF_1 where polyprotein of Sindbis virus was interacting with G3BP1 and G3BP2 proteins in human. Two known interactions for LIG_MYND_1 where E1A protein of Human adenovirus C serotype 5 and EBNA2 protein of Epstein-Barr virus were interacting with ZMYND11 proteins in human. One known interaction was returned for LIG_LYPXL_L_2 where gag protein of Human immunodeficiency virus type 1 was interacting with PDCD6IP protein in human. One known interaction mediated by LIG_Rb_pABgroove_1 where E1A protein of Human adenovirus C serotype 5 had interactions with RB1 protein in human. One known interaction mediated by LIG_TRAF2_2 where LMP1 protein of Epstein-Barr virus had interaction with TRAF2 protein in human. Similarly, one known interaction was returned for LIG_TRAF6 where UL37 protein Human herpesvirus 1 had interaction with TRAF6 protein in human.

Moreover, DMIs were predicted for ELMs including LIG_HCF-1_HBM_1, LIG_KLC1_WD_1,LIG_SH3_2 and TRG_NLS_Bipartite_1. One predicted DMI was mediated by LIG_HCF-1_HBM_1 where UL48 protein of Human herpesvirus 1 was interacting with HCFC1 protein in human, one DMI for LIG_KLC1_WD_1 where VACWR159 protein of Vaccinia virus had interaction with KLC1 protein, one DMI mediated by LIG_SH3_2 where polyprotein of Hepatitis C virus had interaction with GRB2 protein in human. Similarly, one DMI was predicted for TRG_NLS_Bipartite_1 where PB2 protein of Influenza A virus had interaction with the KPNA1 protein in human. All the interacting motifs of viral proteins in predicted DMIs were in the accessible area of the proteins. It should be noted that the predicted interaction might have been reported in literature but are not annotated in ELM.

I further relaxed the DMI-mapping strategy (i.e. ELMc-Domain) **(Figure 2.2)** to predict more DMIs **(Figure 4.5).** This strategy comes at a cost of of higher FDR rate. The total number of predicted DMIs increased 2-3x as compared to the ELMc-Protein strategy, but the overall enrichment score dropped **(Figure 4.6C).** The FDR associated with these predictions was still significant (0.124 in cases of PHISTO and 0.0862 in case of VirHostNet2.0) suggesting that this strategy could still be useful for identifying new DMIs.

It was seen that a few known viral proteins were interacting with multiple different human proteins to hijack host cellular machinery to mediate different functions **(Figure 4.7).** The reason could be the small and complex genome of the viruses which has multifunctional convergently evolved SLiMs. These SLiMs help them mediate number of DMIs to effectively mimic and hijack the host cellular machinery. In general, it can be said that limited genomic resources in the viruses has put intense pressure on them to mediate number of DMIs with their host to maintain their life cycle. According to one study, it was seen that viral proteins are more involved in DMIs, have more SLiMs as compared to human proteins and mimics number of human proteins for their survival (Garamszegi, Franzosa et al. 2013). Looking at the size of our datasets, it was seen that PHISTO was ~1.5x bigger than the VirHostNet2.0 and ~43% of the vhPPIs available in PHISTO were also available in VirHostNet2.0 dataset. As both datasets had shared vhPPIs, it was likely that they also share their predicted DMIs. Approximately, 48% of the predicted DMIs were common in both datasets **(Figure 4.8).**

The ELMc-Domain strategy maintained a modest FDR, suggesting that even noisier DMI predictions might still return a lot of real DMI. To further increase the number of candidate novel DMIs, the mapping stringency was further relaxed to use SLiM occurrences predicted by SLiMProbv2.5.1 (Edwards and Palopoli 2015) instead of known viral instances from ELM. I tried to discover DMIs through both ELMc-Protein and ELMc-domain strategies **(Figure 2.2).** The estimated number of real vhDMI for the SLiMProb-ELMc-Protein strategy was around 145-198 in case of PHISTO, and 89-130 in case of VirHostNet2.0, which indicates that this strategy is predicting real DMIs not identified by the more stringent approaches. However, the FDR of these DMIs was quite high (0.4154 in case of PHISTO and 0.3923 in case of VirHostNet2.0) suggesting that ~40% of predicted DMIs are false positives; individual DMI predictions from this strategy should be interpreted with caution **(Table S 4-1).**

Further relaxing the strategy to use SLiMProb predictions and allow DMI predictions based on interactions between ELM classes and Pfam domain classes (i.e. SLiMProb-ELMc-Domain), substantially increased the numbers of predicted DMIs but dramatically reduced the observed enrichment for predicted SLiM occurrences. Using predicted SLiMs, it should be noted that the estimated false positive rate for individual DMI predictions was very high

(0.7507 in case of PHISTO and 0.7843 in case of VirHostNet2.0). This highlights the need for caution when interpreting naïve large-scale predictions of this nature. In general, implication of both strategies (i.e. ELMc-Protein and ELMc-Domain) using predicted SLiMs generated large number of DMIs but as the FDR was quite high than the known instances, the likelihood of false positive DMIs cannot be ignored. This emphasizes the need to further validate these new predictions to differentiate true positives from false positives **(Table S 4-1).**

4.6.4. How different viral subtypes hijack host cellular machinery through SLiMs

To see how different viruses perturb host cellular machinery through mimicking SLiMs, DMI predictions were made for each viral subtype. The strategy used for this analysis was ELMc-Domain where we used predicted SLiMs. The reason of choosing this strategy over others was the insufficiency of current known DMI/SLiMs knowledge. As this strategy comes with the risk of higher false prediction rate, we removed post translational modifications (PTMs) (*i.e.* MOD and CLV). The reason of excluding these ELM classes was our previous analysis in Chapter 3 where MOD and CLV classes generally looked bad at capturing DMIs **(Figure 3.13)** and were prone to have more random associations. Only four ELM types were included in the analysis (*i.e.* LIG, DEG, DOC and TRG).

RNA viruses can have single stranded RNA or double stranded RNA as their genetic material. These viruses have small genome size (2 to 31 kb) and mostly replicate inside the cytoplasm (Poltronieri, Sun et al. 2015).

dsRNA viruses invade host cells and convert ssRNA to double-stranded genomic RNA. Their genomic dsRNA is then transcribed into mRNA, which upon translation produces proteins essential for viral replication. Eukaryotic systems have defence mechanisms that detect dsRNA and inactivates it through PKR or MDA5 proteins. Therefore, dsRNA viruses replicate their RNA inside icosahedral capsids (Mertens 2004). On the other hand, ssRNA viruses can have a single strand of RNA (either positive or negative), which upon release into the cell helps in replicating the genome (Li, Wei et al. 2013).

Comparison of RNA viruses (i.e. ssRNA and dsRNA) **(Table 4.3)** revealed that ssRNA vhPPI captured more DMIs than dsRNA viruses. The FDR associated with ssRNA predictions was quite high (~0.2-0.7) while dsRNA had lower FDR (~0.1-0.4). Looking at the enrichment trend, dsRNA was more enriched than ssRNA viruses in terms of capturing DMIs. Furthermore, GO enrichment analysis showed that human proteins targeted by ssRNA viral proteins were involved in metabolic and cell regulation related processes. The targeted human proteins were residing in cytoplasmic region of the cell where ssRNA viruses tend to replicate **(Figure 4.13).** On the other hand, human proteins targeted by dsRNA viral proteins were mostly involved in reproduction and metabolic processes. These proteins were mostly residing in plasma membrane **(Figure 4.14).** Different cell cycle regulation and transport related proteins have been reported previously (i.e. P53, ROA2, HNRPK and NPM) which are targeted by RNA viruses (Dyer, Murali et al. 2008; Durmus Tekir, Cakir et al. 2012). The predicted DMI dataset didn't had any of these proteins targeted by RNA viruses. One possible explanation could be that these interactions might not be SLiM mediated or the current SLiM knowledge is not sufficient and more SLiMs/vhPPIs need to be identified.

Just like RNA viruses, DNA viruses can be single stranded (ssDNA) or double stranded (dsDNA). DNA viruses typically have larger genome sizes (10 to 250 kb) as compared to RNA viruses and they replicate inside nucleus of the host cell (Koonin, Krupovic et al. 2015). dsDNA replicates by entering host cell and releasing their DNA into the nucleus. The dsDNA is then transcribed inside the cytoplasm to produce regulatory proteins. These regulatory proteins then help in replication of the dsDNA and transcription of the mRNA. The mRNA is then translated into structural proteins which along with the newly replicated dsDNA are packaged and released outside the cell (Kazlauskas and Venclovas 2011; Rao and Feiss 2015; Kazlauskas, Krupovic et al. 2016). On the other hand, ssDNA viruses first need to convert their ssDNA into dsDNA inside the nucleus which is transcribed into mRNA inside the cytoplasm. This mRNA is translated into different regulatory proteins which help in replication of the genome. The replicated DNA is then again transcribed into mRNA which is translated into structural proteins. The newly replicated ssDNA and the structural proteins are packaged inside a capsid which is then released outside the cell (Krupovic and Forterre 2015).

Comparison of DNA viruses (i.e. dsDNA and ssDNA) revealed that dsDNA captured more DMIs with better enrichment as compared to ssDNA. The FDR associated with both viral subtypes was quite high (~0.4-0.7 for dsDNA and ~0.5-1 for ssDNA) showing that most of the predicted DMIs could be false positives. Gene ontology enrichment analysis showed that human proteins targeted by dsDNA viral proteins were enriched in regulatory and metabolic processes and showed enrichment in nucleus (more specifically nucleoplasm) of the cell. Some of the targeted proteins were also enriched in the cytoplasm **(Figure 4.15).** Similarly, human proteins targeted by ssDNA viral proteins were also found to be enriched in metabolic processes **(Figure 4.16).** If we compare RNA vs DNA viral interactions, RNA viruses were more enriched for DMIs than DNA viruses.

Previous studies have reported that RNA viruses as compared to other viruses are more likely to target proteins which are related to metabolic functions (Pichlmair, Kandasamy et al. 2012). Moreover, in one study the comparison of DNA and RNA vhPPIs revealed that DNA viruses tend to attack cellular and metabolic pathway proteins while RNA viruses tend to attack transport and metabolic proteins (Durmus and Ulgen 2017). This analysis agrees with previous findings in a sense that most of the human proteins targeted by RNA/DNA viruses were primarily involved in metabolic related processes (i.e. protein modification, signal transduction). In general, both DNA viruses and RNA viruses are more likely to attack metabolic proteins through motif mimicry**.**

Overall, looking at the FDR of predictions, it was clear that most of the predicted DMIs could be false positives. Moreover, FDR rate of some viral families was found to be higher in one database than other. The reason could be their curation of data (focusing on particular viruses/samples). This ultimately emphasizes the need of developing more comprehensive vhPPI databases and curating vhPPI data in more effective way.

Looking at the FDR, the likelihood of having false positives was quite high. This was the reason that individual results were not dig into details. Moreover, the high FDR rate of predictions could also have impacted the GO ontology analysis. In general, it can be said that there is a strong need to reduce false discovery rate so that such analysis can become more powerful and reliable. One way to improve such analysis is the availability of more PPI data for under-represented subtypes or to divide viral proteins based on their roles in life cycles. The best thing to do would be to have some filtration steps of predicted DMIs to reduce FDR of predictions. Reducing FDR rate would lead to fewer predictions but will increase higher proportion of real ones. But it was important to first investigate PPIs and find the answer to a broader question "Can viral-human PPIs be a good source of predicting mimicry". Once assured that PPIs are capturing significant real DMIs/DDIs, this knowledge can be used to investigate whether the predicted mimicry candidates are real. For this sort of analysis, it will be important to first reduce FDR and then look for possible true positives. One of the advantages of SLiMEnrich is that it can help in finding DMIs/DDIs through using different sorts of SLiM predictions. For example, SLiMFinder is more tolerant to noise and can be used in conjunction with SLiMEnrich to find DMIs without losing too much signal.

4.6.5. *De-novo* discovery of SLiMs

In network biology, identifying functional SLiMs is considered important as they help in understanding various dynamical process in protein networks. Identifying SLiMs can help in providing clues regarding modes of binding and whether different interactions are likely to be stable or transient. However, computational *de-novo* SLiM prediction is quite challenging, due to their short length and low conservation relative to globular domains. This is the reason high false positive predictions are often anticipated when looking for new SLiM predictions (Prytuliak, Volkmer et al. 2017).

In general, *de-novo* SLiM discovery does not require any prior knowledge of SLiMs to be discovered. It basically works by looking for overrepresented sequence patterns that are unlikely to appear at random. One of the best-performing SLiM discovery tools is SLiMFinder (Edwards, Davey et al. 2007), which uses two dedicated algorithms: SLiMBuild and SLiMChance. SLiMBuild looks for all possible SLiMs based on regular expression in the input dataset with some defined constraints, and SLiMChance assesses the over-representation of the motifs. An alternative version of SLiMFinder known as QSLiMFinder has been developed which uses a query protein sequence to discover SLiMs in a dataset (Palopoli, Lythgow et al. 2015). As QSLiMFinder uses a specific query protein to reduce the motif search space therefore, it increases the sensitivity of the *de-novo* SLiM predictions. QSLiMFinder also reduces the number of datasets returning FP predictions through reducing the number of motifs that could be susceptible to sequence biases in the data (Edwards, Davey et al. 2012; Palopoli, Lythgow et al. 2015). QSLiMFinder is particularly appropriate for prediction of molecular mimicry, where one is specifically interested in sequence patterns in the viral protein.

In this analysis, viral and human interactomes have been integrated to find new motifs using QSLiMFinder. Two random groups were generated as a control to investigate how dataset quality could impact the return of motifs through QSLiMFinder. In control group 1, viral-human interactome was disrupted through shuffling viral proteins (randomvProtein). It was expected that randomising the viral protein would impact the motif search as most of the predicted motifs would be off-targets. In control group 2, the human-human PPI network was disrupted by shuffling human proteins (randomInteractor). This effectively paired each viral protein with a random set of human proteins. As the human-human PPI network was disrupted therefore, a motif needed to be more prevalent (abundant/generic) to be returned. A significant number of datasets in each group returned motifs at SLiMChance P-value ≤ 0.1. This in general implies that the motif prediction was working. The number of datasets returning motifs varied between all three groups where the number of significant datasets for real was higher than control groups at p-value <0.1. This trend was not observed at more stringent p-values (<0.01 and <0.001) where real group didn't look better than control ones **(Figure 4.17A)**. The reason could be that QSLiMFinder might be over-predicting ambiguity and there is a possibility that false positives are dominating the results

The best way to see how good are the predictions and how likely it is returning real motifs is to recover known/true positive (TPs) from the realistic biological data (Edwards, Davey et al. 2012). All the motifs returned from real data (2,564 motifs) were compared with the known ELMs using CompariMotif (Edwards, Davey et al. 2008), and were classified as true positive (TPs) or off-targets (OTs). OTs may represent generic/abundant recurring motifs enriched by chance, or specific motifs that have been enriched in the wrong PPI dataset due to shared interactors (Edwards, Davey et al. 2012). However, these should not be strictly considered as false positives as most of them are highly likely to be real SLiMs, known for biological importance. The number of OTs (generic recurring motifs) being returned from real were 316 while the number of OTs being returned from randomvProtein and randomInteractor group were 217 and 210 respectively. The randomvProtein group and randomInteractor group returned lower OTs than the real group. Both control datasets captured more than 200 OTs. The reason could be that most of the viral proteins might be interacting with same hub protein and randomisation would have essentially linked them with the same hub proteins as in the real group. Another reason could be that the predicted motifs might be over-represented or more prevalent in the network. Looking at the OTs, it was seen that all the predicted OTs were high abundance/generic motifs which are available in multiple proteins. To further see, if the returned motifs were actually real and were known for interactions in ELM, TPs were recovered from all different groups. A motif was regarded as true positive, if and only if, the hub protein was known for interaction in ELM. This analysis returned 12 known interaction motifs in real group, 3 in randomvProtein group and 3 in randomInterctor group **(Figure 4.17B).** The current number of known vhDMIs in databases like ELM is really low and this could be the reason that only few TPs were recovered from the analysis. The randomInteractor and randomvProtein group also returned 3 TPs showing that the true positives in the controls might not actually be true positives themselves. This analysis was based on integration of two datasets (i.e. PHISTO and HI-II-14). The human PPI data used for this analysis was not as big as available comprehensive PPI databases therefore, integration of more comprehensive datasets can certainly improve such analysis through providing more proteins to predict new SLiMs. It has also been shown that masking based on evolutionary conservation can increase the sensitivity of human SLiM prediction and (Davey, Shields et al. 2009) should be explored in the context of viral mimicry.

4.6 Conclusion

High-throughput techniques are being applied to identify vhPPIs and the number of vhPPI data is growing (de Chassey, Meyniel-Schicklin et al. 2014). Despite the progress in identifying vhPPIs, the current DMI knowledge is far from complete. In this study, I have explored vhPPIs as a source of capturing DMIs and found that vhPPI data was capturing DMIs. Both Y2H and AP-MS screens looks promising in terms of identifying DMIs. Looking at the enrichment trend, dsRNA viruses were more enriched than ssRNA viruses in terms of capturing DMIs. On the other hand, ssDNA showed more enrichment than dsDNA viruses. If we compare RNA vs DNA viral interactions, RNA viruses were more enriched in terms of capturing DMIs than DNA viruses. The viral subtype analysis had few limitations for instance, only few vhPPIs were available for dsRNA (58 in case of PHISTO and 165 in case of VirHostNet2.0) and ssDNA (62 in case of PHISTO and 640 in case of VirHostNet2.0) viruses. This low number of vhPPIs could have impacted the overall enrichment of these viruses. Availability of more vhPPI data for these viral subtypes can certainly improve this analysis and can help in predicting more DMIs. I also predicted new SLiMs (2,564 motifs) through integration of human and viral interactomes. Keeping the FDR of DMI predictions in mind, it was likely that most of the predicted DMIs could be false positives. Having a high-throughput in-silico screen for validating individual DMIs and SLiM predictions can be really helpful therefore, the focus of my next chapter was to develop a pipeline for the initial validation of these predictions through *in-silico* structural biology tools. It is highly recommended to reduce the FDR rate before going into experimental validations of these predictions therefore having an initial validation step could be helpful in screening true positives from the false positive interactions.

Supplementary data

Table S 4-1. DMIs predicted using different strategies of SLiMEnrich in PHISTO and VirHostNet2.0 datasets. (https://osf.io/yndsx/?view_only=c035631c7c6b42a38ced7053ddc77799).

5 Chapter 5: *In-silico* structural evaluation of SLiMs and DMIs

5.1 Abstract

Domain-motif interactions (DMIs) are transient interactions which occur when a Short Linear Motif (SLiM) binds with a globular domain through a small contact interface. To understand how DMIs occur to maintain different regulatory processes and to see how different viruses hijack host cellular machinery, it is crucial to have knowledge of the binding mode of these DMIs. The degenerate nature and small contact interface make it difficult to identify DMIs through traditional *in-vitro* and *in-vivo* experiments. As the predictions come with high false positive rate, there is a need for high-throughput *in-silico* validations of these predictions before going into experimental work. Here, I have evaluated binding energy changes of predicted SLiM instances through *in-silico* peptide exchange experiments to see how they bind with the known 3D DMI complexes. Only a few of the predicted mimicry candidates from previous analysis **(Chapter 4)** showed binding with the native DMI structures. The analysis done in this chapter was a pilot study and further validation through additional computational as well as experimental techniques is much needed.

5.2 Background

During recent years, interactome maps of several organisms have been drafted. These interactome maps are being used to unveil protein interactions in a high-throughput manner (Rual, Venkatesan et al. 2005). However, high-throughput methods are not enough to find clues about underlying molecular details of protein interactions. Atomic level investigation is often required to see how two proteins interact with each other and to find which residues are in contact between two proteins. Currently, this is only possible through resolving three-dimensional structures to characterize interfaces involved in interactions. The 3D structural information available in Protein Data Bank (PDB) can be used to characterize protein interactions based on their contact interfaces (i.e. domain-domain interaction and domain-motif interaction) (Aloy and Russell 2006). Domain-Domain Interactions create large contact interfaces (2.000Å2) between two globular domains. On the other hand, SLiMs in DMIs establish small contact interface with their interacting domain partners, which makes it challenging to achieve a high prediction specificity. This is the reason that new computational methods are required to validate them before going *in-vitro* (Stein and Aloy 2008).

Despite the improvements in PPI detection experiments, the current DMI number is still underrepresented (Pawson and Linding 2005) and only a small fraction of known DMIs is available in databases like ELM (Dinkel, Van Roey et al. 2016) and 3did (Mosca, Ceol et al. 2014). DMIs are found to be highly specific in *in-vivo* experiments where a SLiM binds with globular domain of a specific protein only (e.g. DMI between Pbs2 peptide and SH3 domain of Sho1)(Zarrinpar, Park et al. 2003). The atomic contacts/bonds between these interactions are insufficient to explain this high degree of specificity. Thus, just like phosphorylation events, biological context (i.e. subcellular localization, expression patterns) is what determines the interaction specificity in DMIs (Linding, Jensen et al. 2007). An example of such case is GYF domain of CD2BP2 protein which is localized in soluble membrane region of protein and doesn't compete with SH3 domain of Fyn as it is in lipid rafts in T-cells. In *in-vivo* experiments this interaction is impossible, but this can happen in *in-vitro* experiments (Freund, Kuhne et al. 2002).

In Chapter 4, viral SLiMs and DMIs were predicted using different strategies and some of them could potentially be real. To see if the predictions were real, a pilot study was designed for the initial validation of the predictions. Therefore, the main focus of this chapter is to determine whether binding specificity can help in initial screening of DMIs before going *in-vitro*. Here, known DMI data from 3did (Mosca, Ceol et al. 2014) and ELM (Dinkel, Van Roey et al. 2016) is combined with *in silico* peptide exchange experiments (Kiel and Serrano 2014) to see whether predicted changes in binding energies for known/predicted motifs can be used to discriminate real motif occurrences from non-binding peptide sequences. If worked successfully, the outcome of this analysis can help differentiate real interaction motifs from false positives and will be useful for initial validation of predicted DMIs.

5.3 Aims and Objectives

This chapter aims to establish whether changes in the predicted binding affinity of SLiMs with globular domains can be used to discriminate real motif occurrences from non-binding peptide sequences.

The specific objectives were:

- To see if changes in predicted binding energies can be used to discriminate real motif occurrences from non-binding peptide sequences.
- To quantify predicted binding energies of predicted viral instances through *in silico* peptide exchange experiment.
- To see if predicted DMIs can be validated through quantification of binding energies.

5.4 Methods

5.4.1 3D DMI data retrieval

Known domain-motif interaction (DMI) data was downloaded from 3did (Mosca, Ceol et al. 2014) on 2018-11-01 by retrieving DMIs that had solved PDB structures. The 3D DMI data from 3did was cross-referenced with ELM (Dinkel, Van Roey et al. 2016) to get motif information (i.e. motif sequence and positions). 3D structures of known DMIs were retrieved from PDB (Rose, Prlic et al. 2017) on 2018-11-08. A total of 100 structures belonging to 47 different ELM types, which accounts to 130 different DMIs, were selected. Foldx does not work with any non-standard residues (including post-translational modifications), therefore structures with any non-standard residues in the bound ligand were excluded from the analysis. Out of 100 3D structures, only 23 were selected for further analysis**.** The selected 23 DMI complexes belonged to 12 different ELMs and included degron, docking, targeting and ligand motifs.

5.4.2 SLiM prediction

Viral proteins in previously downloaded datasets: PHISTO (Durmus Tekir, Cakir et al. 2013) and VirHostNet2.0 (Guirimand, Delmotte et al. 2015) **(Chapter 4: 4.4.1)** were used to predict new instances of known ELMs through SLiMProb v2.5.1 (Edwards and Palopoli 2015) tool **(Chapter 4: 4.4.2)**. Similarly, reviewed proteins of human proteome were retrieved from Uniprot (Apweiler, Bairoch et al. 2004) on 2018-07-20 and were used to predict SLiMs using SLiMProb v2.5.1 (Edwards and Palopoli 2015) with disordered masking (IUPred score >= 0.2) (Hagai, Azia et al. 2011) and no evolutionary filtering of results.

5.4.3 Peptide dataset generation

Four peptide datasets (one test and three control) for each known 3D DMI were generated: **1)** Viral (test) dataset which had predicted SLiMs of viral proteome **(Chapter 4: 4.4.2)**, **2)** True positive (control) dataset which had known motif instances (experimentally validated) from ELM for the same ELM class, **3)** Human (control) dataset which had predicted SLiMs from the human proteome (500 random occurrences if number of SLiM occurrences was >500), and **4)** True negative (control) dataset which had 500 randomly generated peptide sequences, using a uniform frequency distribution of all twenty amino acids **(Figure 5.1).**

5.4.4 Structural optimization

FoldX (Schymkowitz, Borg et al. 2005) was used for the structural optimization of the known DMI complexes. FoldX provides a fast and quantitative estimation of the importance of the interactions contributing to the stability of proteins and protein complexes. A short optimization of the structure of the native DMI complex was performed to eliminate small clashes and other undesirable features by the RepairPDB function within FoldX. The RepairPDB command helps in repairing residues which have bad torsion angles, van der Waals' clashes, or total energy (option VdWDesign=2). Moreover, water and all non-protein ligands were removed from the structures. Gibbs free energy (ΔG) of known complexes was calculated using "Stability" command of FoldX to assess global folding stability of the complex. ΔG gives idea of the stability of a protein, which is expressed in kcal/mol (Lower ΔG = More Stability).

5.4.5 FoldX model construction and binding energy calculations

New DMI complexes were generated using optimized known structures by swapping already bound peptide with different peptide datasets using BuildModel command of FoldX and binding energy changes were calculated for each peptide. FoldX works by first mutating each amino acid to alanine and then annotating the side-chain energies of the neighbouring residues. It then mutates the alanine to selected amino acid and recalculates the side-chain energies of neighbouring residues.

Binding energy difference (ΔΔG) between known and new DMI complex was calculated by:

$$\Delta\Delta G = \Delta G_{\text{mutant}} - \Delta G_{\text{wildtype}}$$

where ΔG_{mutant} and $\Delta G_{\text{wildtype}}$ are the binding free energies of the mutant complex (new DMI complex) and the wild-type complex (known DMI complex), respectively. ΔΔG was calculated for all peptides in each dataset (known, viral, human and true negative) by swapping the already bound peptide with different peptides **(Figure 5.1).** It was manually confirmed that motifs were being swapped in right positions through manually confirming the list of SLiMs (mutant file) required by the FoldX.

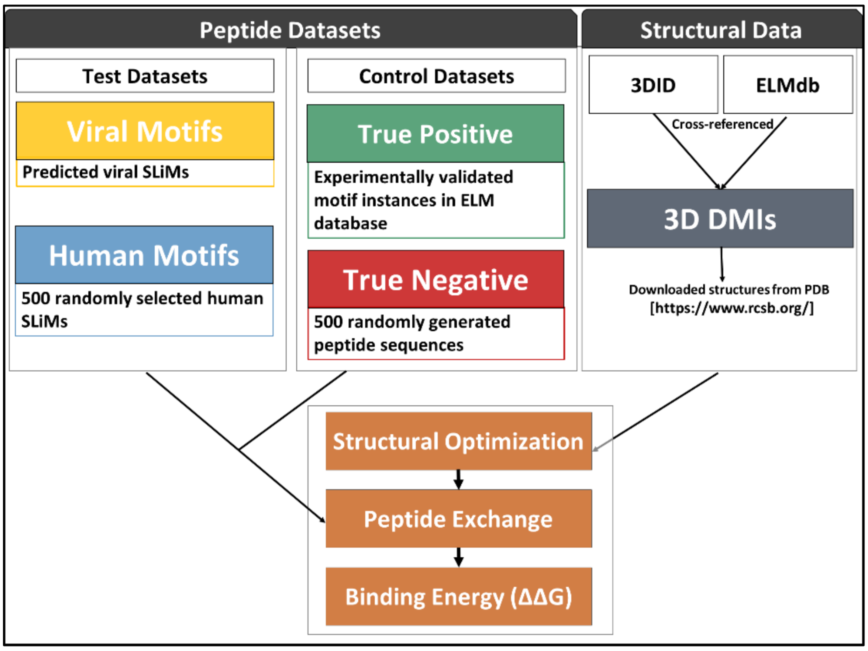


Figure 5.1. Workflow of evaluating binding vs non-binding SLiMs.

3D DMI data was retrieved from ELM and 3did databases. 3D structures were retrieved from PDB and were optimized using FoldX. Two control datasets were generated: true positive dataset which had known SLiMs and true negative dataset which had randomly generated peptides. Two validation datasets were generated: Viral motifs which had predicted viral SLiMs and human motifs which had 500 randomly selected predicted human SLiMs. All peptide datasets were exchanged with the already bound peptide in the complex and binding energy (ΔΔG) was calculated.

5.4.6 Statistical analysis

A two-tailed non-parametric Wilcoxon rank test was performed to detect significant differences in ΔΔG distributions. For this purpose, four different comparisons were done:

1) TP vs TN: This was done to see if there will be a signal.
2) Human vs True Negative: This was done to see if there was any recognisable binding motif.
3) TP vs Human: This was done find if true positive SLiMs were binding more strongly that the random human occurrences.
4) Virus vs Human: This was done to see whether viral SLiMs were binding strongly than the random human ones.

5.5 Results

5.5.1 Data retrieval and structural optimization

First, all known 3D DMI structures were optimised using the repair function of FoldX. During this procedure, residues are identified that have bad total energies or van der Waal's clashes; they are self-mutated and replaced by another, more favourable rotamer or subrotamer. Then, the Gibbs free energy (ΔG) of the complex was calculated to see how stable the known complex was **(Table 5.1).**

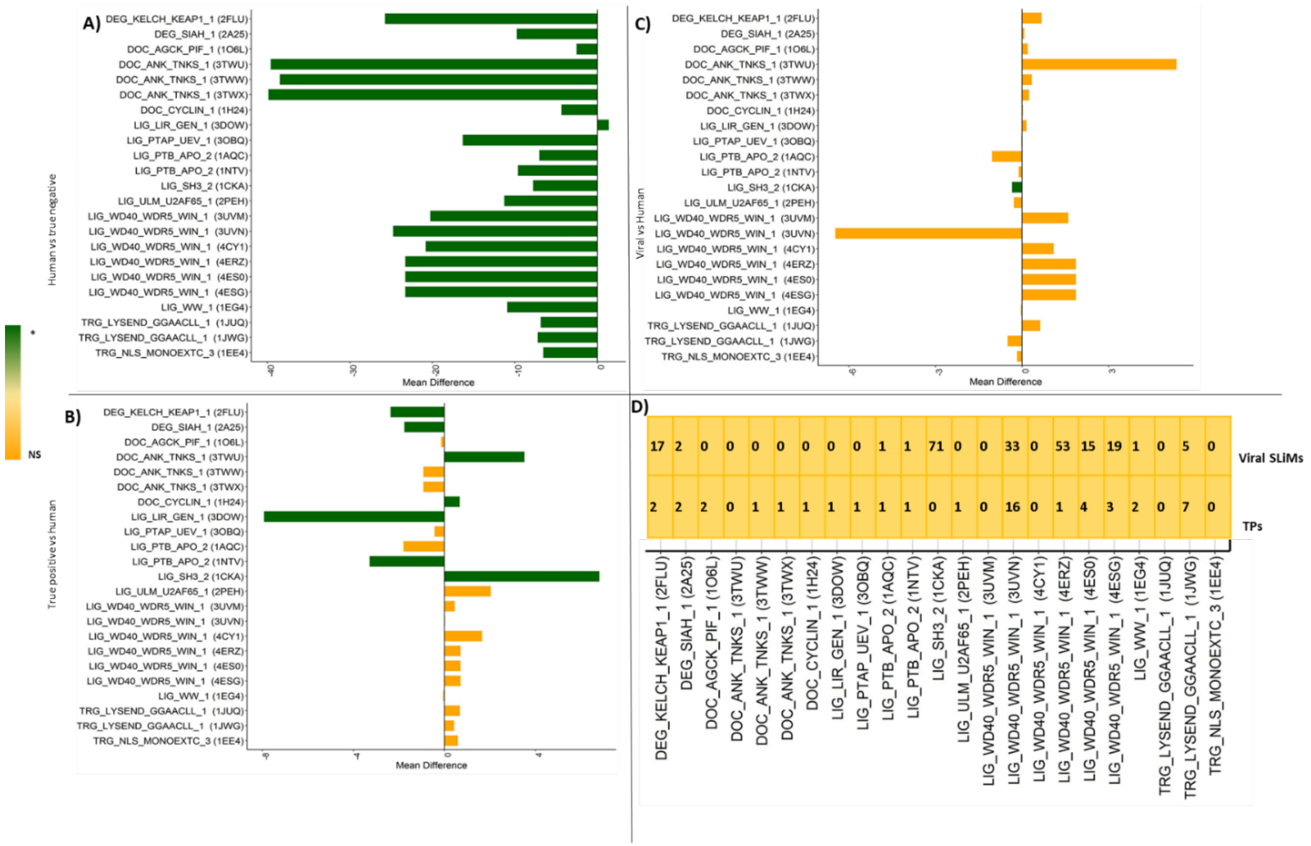
Table 5.1. Data statistics and binding energies (ΔG) of known DMI complexes.

ELM Group	ELM Type	PDB ID	Motif Sequence	Motif Size	Full Peptide Sequence	Resolution (Å)	Binding affinity (ΔG) kcal/mol	Number of viral SLiMs	Number of TPs
Degron Motifs	DEG_KELCH_KEAP1_1	2FLU	DEETGE	6	AFFAQLQLDEETGEFL	1.5 Å	-30.52	15	9
	DEG_SIAH_1	2A25	PAAVVAP	7	EKPAAVVAPITTG	2.2 Å	-48.60	10	9
Docking Motifs	DOC_AGCK_PIF_1	1O6L	RTTSF	5	GRPRTTSFAE	1.6 Å	-44.77	2	8
	DOC_ANK_TNKS_1	3TWU	RPPPIG	6	SRRVARPPPIGAEVPN	1.8 Å	-36.89	111	17
		3TWW	RQSPDG	6	LPHLQRQSPDGQSFRS	2 Å	-35.00		
		3TWX	RESPDG	6	LPHLQRESPDGQSFRS	1.8 Å	-56.31		
	DOC_CYCLIN_1	1H24	RRL	3	PVKRRLDLE	2.5 Å	-5.41	203	28
Ligand Motifs	LIG_LIR_GEN_1	3DOW	WDFL	4	SLEDDWDFLPPX	2.3 Å	-17.26	533	19
	LIG_PTAP_UEV_1	3OBQ	PSAP	4	PTPSAPVPL	1.4 Å	-30.29	48	18
	LIG_PTB_APO_2	1AQC	ENPTY	5	GYENPTYKFF	2.3 Å	-23.26	151	17
		1NTV	DNPVY	5	NFDNPVYRKT	1.5 Å	-23.84		
	LIG_SH3_2	1CKA	PPALPPK	6	PPPALPPKKR	1.5 Å	-14.58	239	16
	LIG_ULM_U2AF65_1	2PEH	SRWDE, KSRWD	5	KRKSRWDETP	2.11 Å	-17.35	3	5
	LIG_WD40_WDR5_WIN_1	3UVM	ARAE	4	GAARAEVYLR	1.57 Å	-15.71	2	5
		3UVN	ARSE	4	GSARSEGYYPI	1.792 Å	-9.84		
		4CY1	ARTR	4	DGTCVAARTRPVLSY	1.5 Å	-1.08		
		4ERZ	ARAE	4	LNPHGAARAEVYLR	1.75 Å	-17.90		
		4ES0	ARSE	4	EHVTGCARSEGFYT	1.817 Å	-3.16		
		4ESG	ARAE	4	EPPLNPHGSARAEVHLR	1.7 Å	-7.19		
	LIG_WW_1	1EG4	PPPY	4	KNMTPYRSPPPYVPP	2 Å	-45.12	15	7
Targeting Motifs	TRG_LYSEND_GGAACLL_1	1JUQ	DDHLL	5	EESEERDDHLLPM	2.2 Å	-113.28	2	5
		1JWG	DEDLL, DLLHI	5	SFHDDSDEDLLHI	2 Å	-54.68		
	TRG_NLS_MONOEXTC_3	1EE4	KRVK	4	PAAKRVKLD	2.1 Å	-178.92	225	18

5.5.2 Peptide exchange experiment

Optimized 3D complexes were used to evaluate energy differences of predicted SLiMs to distinguish between binding vs non-binding motifs. For this purpose, peptide datasets were swapped with the already bound peptide sequence. As already bound peptides were longer than the actual motif sequence (flanking residues accompanying the actual motif sequences), only the motif sequence was swapped by selecting the same length SLiMs from the peptide datasets. Foldx peptide exchange works by giving a file (mutant file) having a reference sequence (sequence to be exchanged) along with a list of peptides that are to be exchanged. Only the motif sequence (in some cases a few flanking residues as well) was given as reference peptide sequence while keeping the flanking residues unchanged. Binding energy differences were calculated for all viral SLiMs, randomly selected human SLiMs (selected 500 if number of SLiMs was quite large), all known instances in ELM and 500 randomly generated peptide sequences. The known instances (true positives) and random occurrences of human SLiMs were used to find the answer of our general question whether binding energy differences can help in differentiating good vs bad binders and then to check whether this knowledge could be used to find viral SLiMs that could be mimicry candidates. To get the general idea of how viral SLiMs were binding with the native structure, I first evaluated binding energy differences of all viral SLiM instances and then looked for any mimicry candidates among those SLiMs. The mimicry candidate data was retrieved from the predicted DMI dataset using the ELMc-domain strategy **(Chapter 4: 4.4.2).** The predicted viral SLiMs were found to be stabilising for 10 ELM classes (i.e. DEG_SIAH_1, DOC_ANK_TNKS_1, DOC_CYCLIN_1, LIG_PTAP_UEV_1, LIG_PTB_APO_2, LIG_SH3_2, LIG_ULM_U2AF65_1, LIG_WW_1, TRG_LYSEND_GGAACLL_1 and TRG_NLS_MONOEXTC_3). Out of 23 analysed structures, 47% demonstrated stabilisation ($\Delta\Delta G < 0$) with the predicted viral peptides **(Figure 5.2D)**. All ELMs showed significant stabilisation for human occurrences in comparison to true negatives **(Figure 5.2A)**. A total of 4 structures (i.e. DEG_KELCH_KEAP1_1 (2FLU), DEG_SIAH_1 (2A25), LIG_LIR_GEN_1

(3DOW) and LIG_PTB_APO_2 (1NTV)) showed significant stabilisation for true positives in comparison to random occurrences **(Figure 5.2B)** while only 1 structure (LIG_SH3_2 (1CKA)) showed significant stabilisation for viral SLiMs as compared to random human occurrences. **(Figure 5.2C)**. In general, in most of the cases, true positives were not distinguishable from random human occurrences and viral SLiMs were not found to be distinguishable from the human occurrences.



	DEG_KELCH_KEAP1_1 (2FLU)	DEG_SIAH_1 (2A25)	DOC_AGCK_PIF_1 (1O6L)	DOC_ANK_TNKS_1 (3TWU)	DOC_ANK_TNKS_1 (3TWW)	DOC_ANK_TNKS_1 (3TWX)	DOC_CYCLIN_1 (1H24)	LIG_LIR_GEN_1 (3DOW)	LIG_PTAP_UEV_1 (3OBQ)	LIG_PTB_APO_2 (1AQC)	LIG_PTB_APO_2 (1NTV)	LIG_SH3_2 (1CKA)	LIG_ULM_U2AF65_1 (2PEH)	LIG_WD40_WDR5_WIN_1 (3UVM)	LIG_WD40_WDR5_WIN_1 (3UVN)	LIG_WD40_WDR5_WIN_1 (4CY1)	LIG_WD40_WDR5_WIN_1 (4ERZ)	LIG_WD40_WDR5_WIN_1 (4ES0)	LIG_WD40_WDR5_WIN_1 (4ESG)	LIG_WW_1 (1EG4)	TRG_LYSEND_GGAACLL_1 (1JUQ)	TRG_LYSEND_GGAACLL_1 (1JWG)	TRG_NLS_MONOEXTC_3 (1EE4)
Viral SLiMs	17	2	0	0	0	0	0	0	0	1	1	71	0	0	33	0	53	15	19	1	0	5	0
TPs	2	2	2	0	1	1	1	1	1	1	1	0	1	0	16	0	1	4	3	2	0	7	0

Figure 5.2. Wilcoxon rank test of binding energy differences of different peptide datasets.

A) Human vs True negative comparison. X-axis shows mean binding energy difference between human and TNs. Green indicates significant p-value (P-value < 0.05), and yellow indicates non-significant p-values (p-value > 0.05), **B)** True positive vs Human comparison. X-axis shows mean binding energy difference between human and TNs. Green indicates significant p-value (P-value < 0.05), and yellow indicates non-significant p-values (p-value > 0.05), **C)** Viral vs Human comparison. X-axis shows mean binding energy difference between human and TNs. Green indicates significant p-value (P-value < 0.05), and yellow indicates non-significant p-values (p-value > 0.05), **D)** Number of stabilising true positive and viral SLiM occurrences for each analysed PDB structure.

5.5.2.1 Degron Motifs

In this analysis, two degron ELMs (*i.e.* DEG_KELCH_KEAP1_1 and DEG_SIAH_1) have been evaluated to find binding vs non-binding SLiMs.

1- **DEG_KELCH_KEAP1_1**

The DEG_KELCH_KEAP1_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon p < 0.05) **(Figure 5.3, Figure 5.2A).** True positive SLiMs showed less destabilisation than true negatives (Wilcoxon p < 0.05). All human and true positive SLiM occurrences were predicted to destabilise the structure (PDB ID: 2FLU, ΔΔG > 0), with the true positives demonstrating less destabilisation than random human occurrences (Wilcoxon p < 0.05) **(Figure 5.2B)**. Viral SLiM were also destabilising (ΔΔG > 0) and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**.

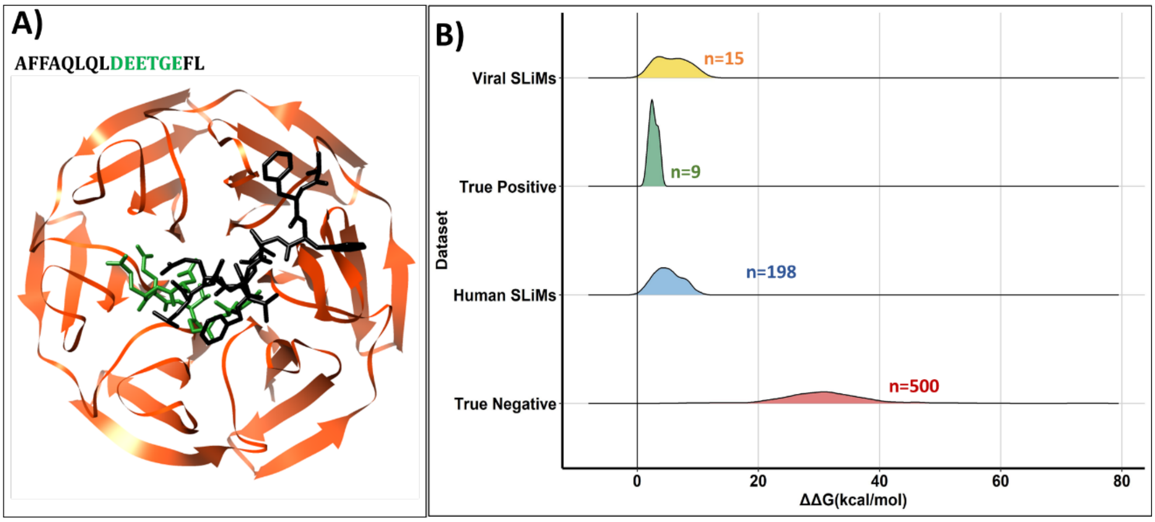


Figure 5.3. Peptide exchange experiment using known 3D DMI (PDB ID: 2FLU) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of DEG_KELCH_KEAP1_1 DMI complex (PDB ID: 2FLU), **B)** Binding energy differences (ΔΔG) of different peptide datasets bound to native DEG_KELCH_KEAP1_1 complex.

2- DEG_SIAH_1

The DEG_SIAH_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon p < 0.05) **(Figure 5.4, Figure 5.2A).** True positive SLiMs showed higher stabilisation than true negatives (Wilcoxon p < 0.05). Most of the human and true positive SLiM occurrences were predicted to stabilise the structure (PDB ID: 2A25, ΔΔG < 0), with the true positives demonstrating higher stabilisation than random human occurrences (Wilcoxon p < 0.05) **(Figure 5.2B)**. A total of 5 viral SLiMs were also stabilising (ΔΔG < 0) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**.

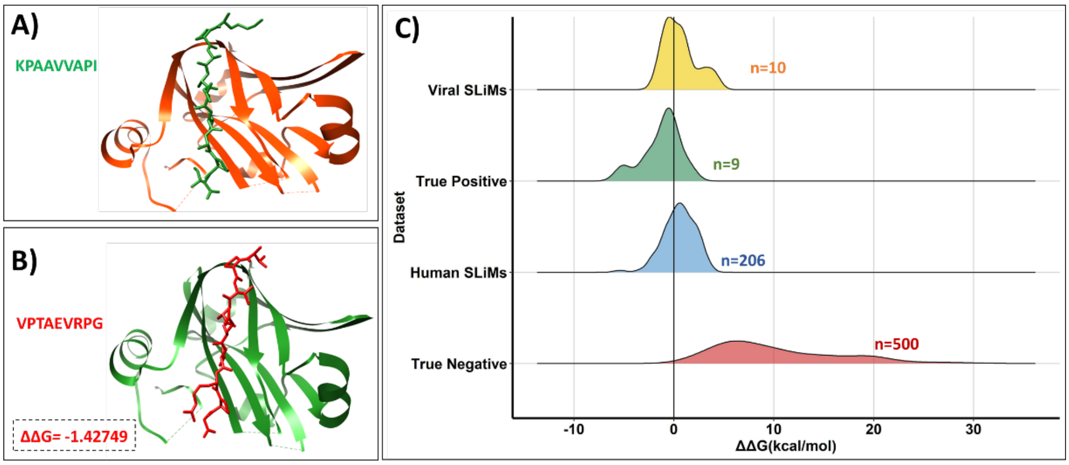


Figure 5.4. Peptide exchange experiment using known 3D DMI (PDB ID: 2A25) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native DEG_SIAH_1 complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native DEG_SIAH_1 complex.

5.5.3 Docking Motifs

Docking motifs are SLiMs present within the substrates which are removed from the phosphorylation sites which promote high affinity interactions with kinases through interactions with outside regions of the catalytic site of the enzymes (Lee, Hoofnagle et al. 2004). 3D complexes of DMIs of three docking motifs (i.e. DOC_AGCK_PIF_1, DOC_ANK_TNKS_1 and DOC_CYCLIN_1) were used to evaluate binding affinities of predicted SLiMs.

1- DOC_AGCK_PIF_1

The DOC_AGCK_PIF_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon p < 0.05) **(Figure 5.5, Figure 5.2A).** True positive SLiMs showed less destabilisation than true negatives (Wilcoxon p < 0.05). All human and true positive SLiM occurrences were predicted to destabilise the structure (PDB ID: 1O6L, ΔΔG > 0). The true positives could not be distinguished from random human occurrences (Wilcoxon p > 0.05) **(Figure 5.2B)**. Viral SLiMs were also destabilising (ΔΔG > 0) and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**.

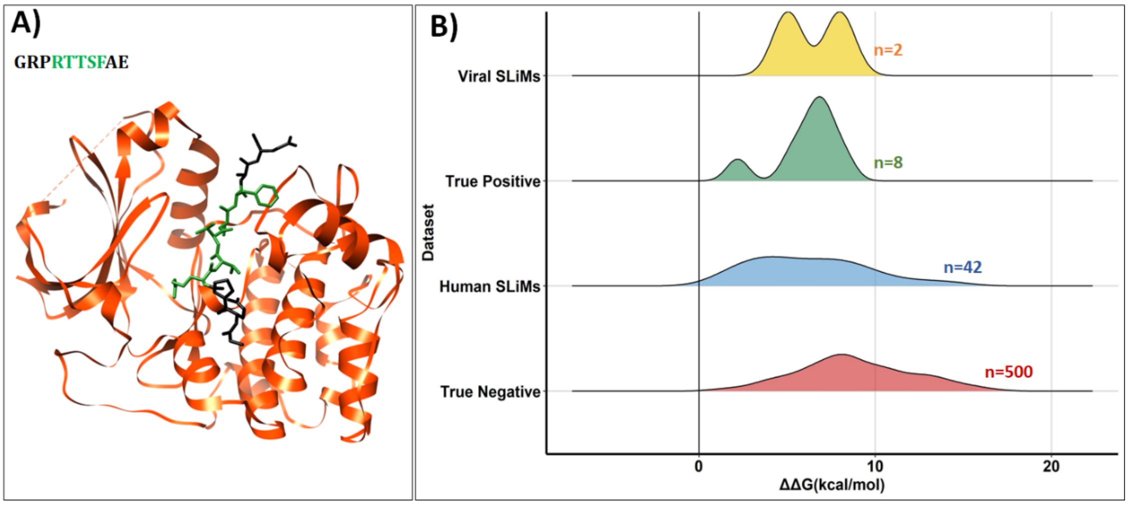


Figure 5.5. Peptide exchange experiment using known 3D DMI (PDB ID:1O6L) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of the known DOC_AGCK_PIF_1 complex, **B)** Binding energy differences (ΔΔG) of different datasets bound to native DOC_AGCK_PIF_1 complex.

2- DOC_ANK_TNKS_1

DOC_ANK_TNKS_1 had three known complexes (i.e. 3TWU, 3TWW and 3TWX). True negative vs true positive comparison for all three complexes demonstrated that true positives were more stabilising than the true negatives (Wilcoxon $p < 0.05$). The DOC_ANK_TNKS_1 (PDB ID: 3TWU) motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.6, Figure 5.2A).** Only 2 human and 2 true positive SLiM occurrences were predicted to stabilise the structure (PDB ID: 3TWU, $\Delta\Delta G < 0$), with the true positives demonstrating less destabilisation than random human occurrences (Wilcoxon $p < 0.05$) **(Figure 5.2B)**. Only one viral SLiM was stabilising ($\Delta\Delta G < 0$) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2C)**. None of the mimicry candidates showed effective binding with the native complex.

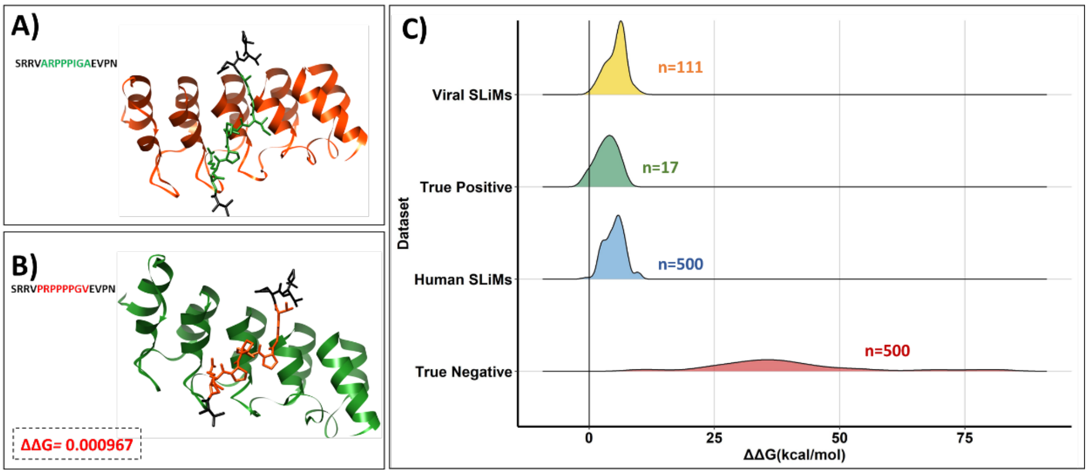


Figure 5.6. Peptide exchange experiment using known 3D DMI (PDB ID: 3TWU) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native DOC_ANK_TNKS_1 (3TWU) complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native DOC_ANK_TNKS_1 (3TWU) complex.

The DOC_ANK_TNKS_1(PDB ID: 3TWW) motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.7, Figure 5.2A).** Only 93 human and 3 true positive SLiM occurrences were predicted to stabilise the structure (PDB ID: 3TWW, $\Delta\Delta G < 0$). The true positives could not be distinguished from random human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2B)**. A total of 19 viral SLiM were predicted to be stabilising ($\Delta\Delta G < 0$) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2C)**. Only one mimicry candidate from predicted DMI dataset (TRPGPPGI) was found to be stabilising ($\Delta\Delta G < 0$).

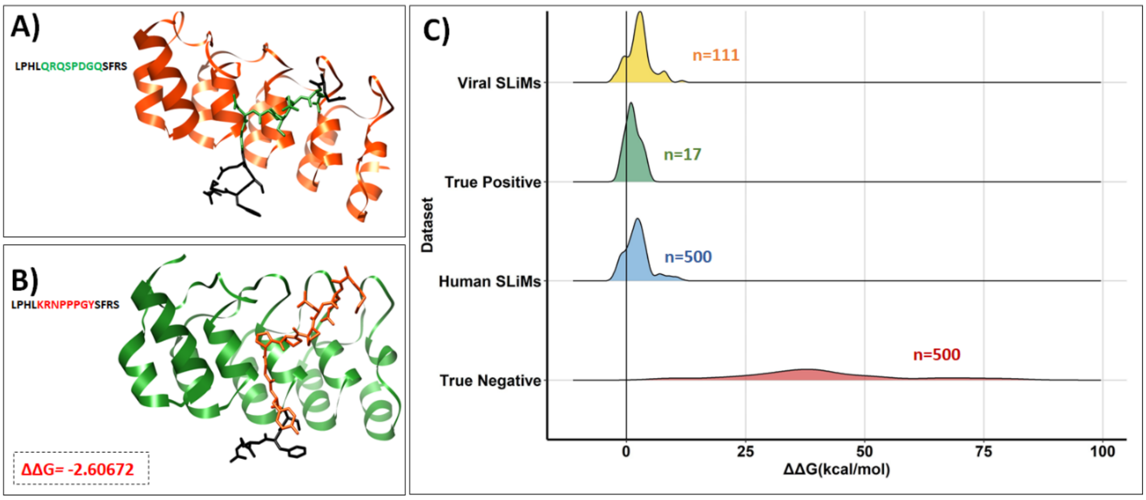


Figure 5.7. Peptide exchange experiment using known 3D DMI (PDB ID: 3TWW) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native DOC_ANK_TNKS_1 (3TWW) complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native DOC_ANK_TNKS_1 (3TWW) complex.

The DOC_ANK_TNKS_1 (PDB ID: 3TWX) motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.8, Figure 5.2A).** Only 70 human and 4 true positive SLiM occurrences were predicted to stabilise the structure (PDB ID: 3TWX, $\Delta\Delta G < 0$). The true positives could not be distinguished from random human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2B)**. A total of 15 viral SLiM were predicted to be stabilising ($\Delta\Delta G < 0$) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2C)**. None of the mimicry candidate from predicted DMI dataset was found to be stabilising.

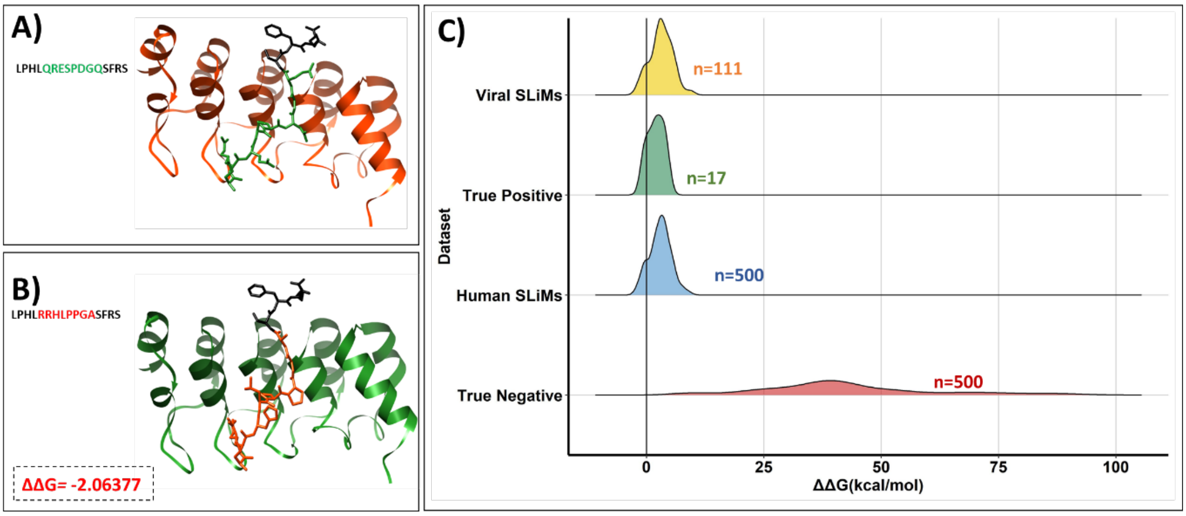


Figure 5.8. Peptide exchange experiment using known 3D DMI (PDB ID: 3TWX) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native DOC_ANK_TNKS_1 (3TWX) complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native DOC_ANK_TNKS_1 (3TWX) complex.

3- DOC_CYCLIN_1

The DOC_CYCLIN_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.9, Figure 5.2A).** True positive SLiMs showed less destabilisation than true negatives (Wilcoxon $p < 0.05$). A total of 100 human and only 1 true positive SLiM occurrence were predicted to stabilise the structure (PDB ID: 1H24, $\Delta\Delta G < 0$), with the true positives demonstrating less destabilisation than random human occurrences (Wilcoxon $p < 0.05$) **(Figure 5.2B).** A total of 53 viral SLiM were found to be stabilising ($\Delta\Delta G < 0$) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2C).**

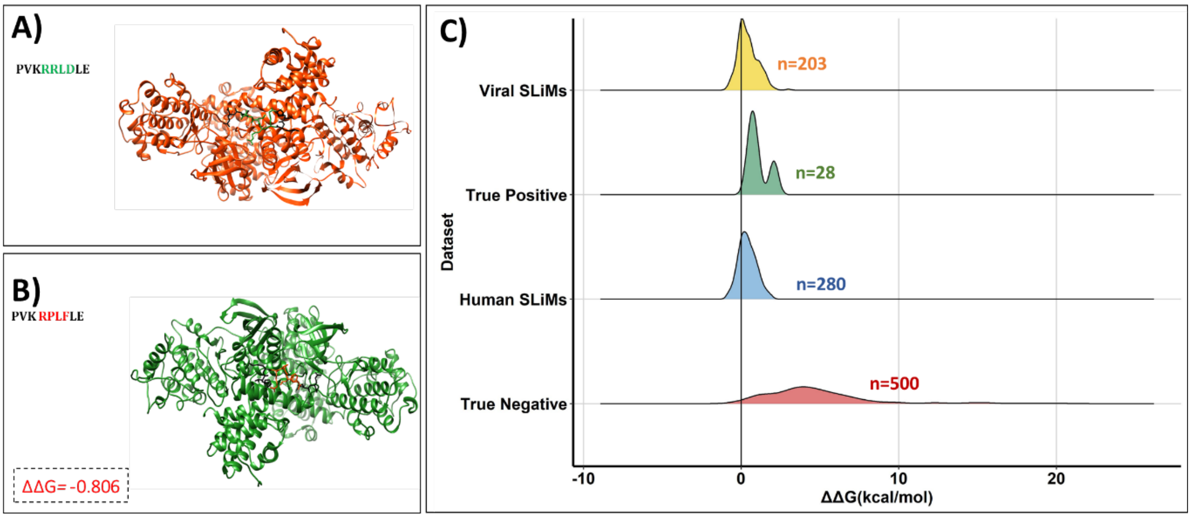


Figure 5.9. Peptide exchange experiment using known 3D DMI (PDB ID: 1H24) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native DOC_CYCLIN_1 complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native DOC_CYCLIN_1 complex.

5.5.4 Ligand binding sites

Ligand binding sites have crucial role in carrying out various biochemical functions of proteins. Ligands are usually small molecules which produce signals upon binding with specific sites in target proteins (Kinoshita and Nakamura 2005). 3D DMI complexes of 7 ligand motifs were selected (i.e. LIG_LIR_GEN_1, LIG_PTAP_UEV_1, LIG_PTB_APO_2, LIG_SH3_2, LIG_ULM_U2AF65_1, LIG_WD40_WDR5_WIN_1 and LIG_WW_1) **(Table 5.1)**.

1- LIG_LIR_GEN_1

The LIG_LIR_GEN_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.10, Figure 5.2A).** True positive SLiMs showed less destabilisation than true negatives (Wilcoxon $p < 0.05$). All human and true positive SLiM occurrences were predicted to destabilise the structure (PDB ID: 3DOW, ΔΔG > 0), with the true positives demonstrating more stabilisation than random human occurrences (Wilcoxon $p < 0.05$) **(Figure 5.2B)**. Viral SLiM were also destabilising

(ΔΔG > 0) and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**.

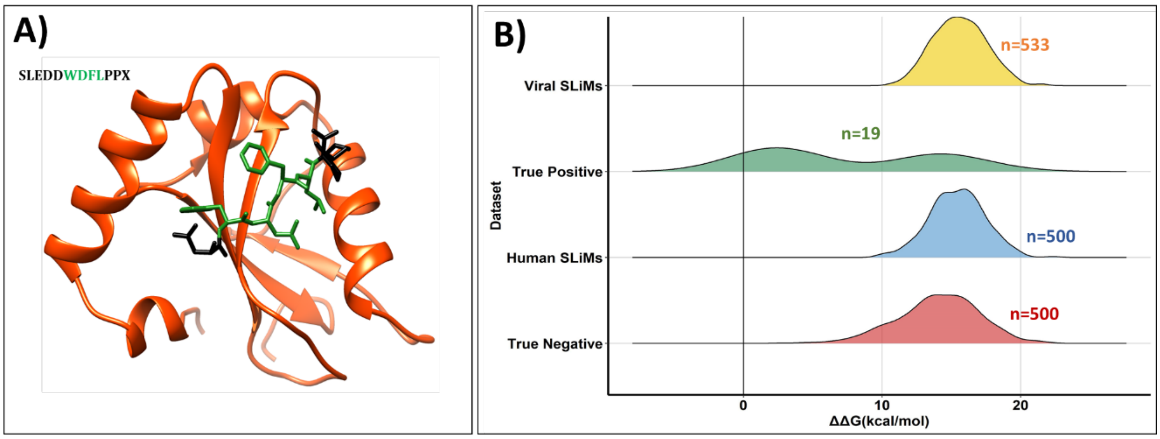


Figure 5.10. Peptide exchange experiment using known 3D DMI (PDB ID: 3DOW) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of LIG_LIR_GEN_1 DMI complex (PDB ID: 3DOW), **B)** Binding energy differences (ΔΔG) of different peptide datasets bound to native LIG_LIR_GEN_1 complex.

2- LIG_PTAP_UEV_1

The LIG_PTAP_UEV_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon p < 0.05) **(Figure 5.11, Figure 5.2A).** True positive SLiMs showed less destabilisation than true negatives (Wilcoxon p < 0.05). Most of the human and true positive SLiM occurrences were predicted to stabilise the structure (PDB ID: 3OBQ, ΔΔG < 0), with the true positives demonstrating less destabilisation than random human occurrences (Wilcoxon p < 0.05) **(Figure 5.2B)**. A total of 33 viral SLiMs were found to be stabilising (ΔΔG < 0) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**. None of the mimicry candidates was found to be stabilising.

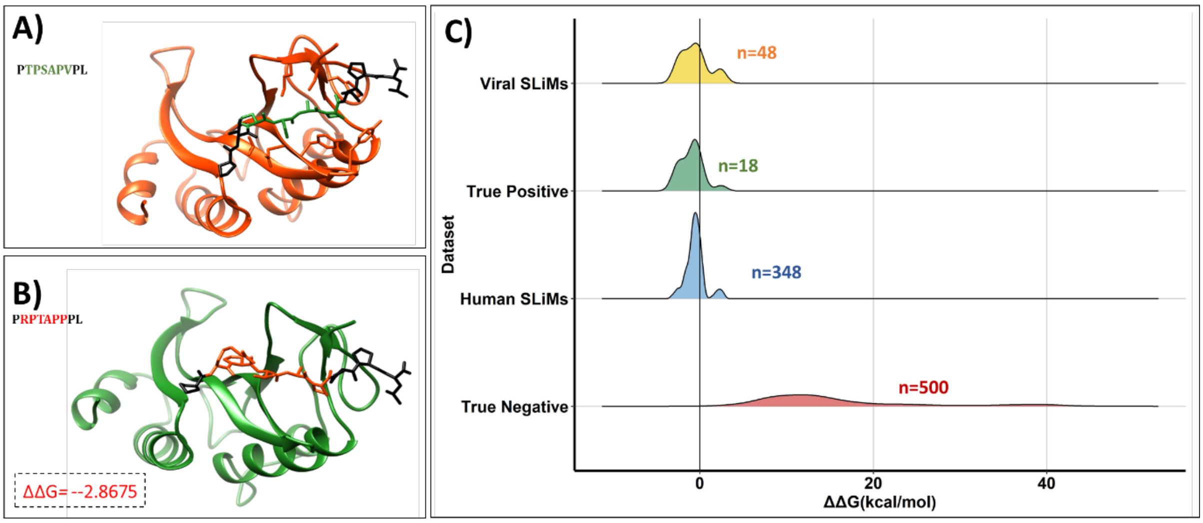


Figure 5.11. Peptide exchange experiment using known 3D DMI (PDB ID: 3OBQ) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native LIG_PTAP_UEV_1 complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native LIG_PTAP_UEV_1 complex.

3- LIG_PTB_APO_2

LIG_PTB_APO_2 had 2 solved DMI structures (i.e. 1NTV and 1AQC). The LIG_PTAP_UEV_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon p < 0.05) **(Figure 5.12, Figure 5.2A).** All the human SLiM occurrences were predicted to destabilise the structure (PDB ID: 1NTV, ΔΔG > 0) for both structures (PDB ID: 1NTV and 1AQC) and only 1 true positive SLiM occurrence was found to be stabilising (ΔΔG < 0) with the PDB structure 1NTV while all the true positive SLiM occurrence were destabilising the 1AQC structure (ΔΔG > 0). In case of 1AQC structure, true positives were demonstrating less destabilisation than random human occurrences (Wilcoxon p < 0.05) while in case of 1NTV, true positives were demonstrating more stabilisation than random human occurrences (Wilcoxon p < 0.05) (**Figure 5.2B)**. All the viral SLiM occurrences were found to be destabilising (ΔΔG > 0) for both structures and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**.

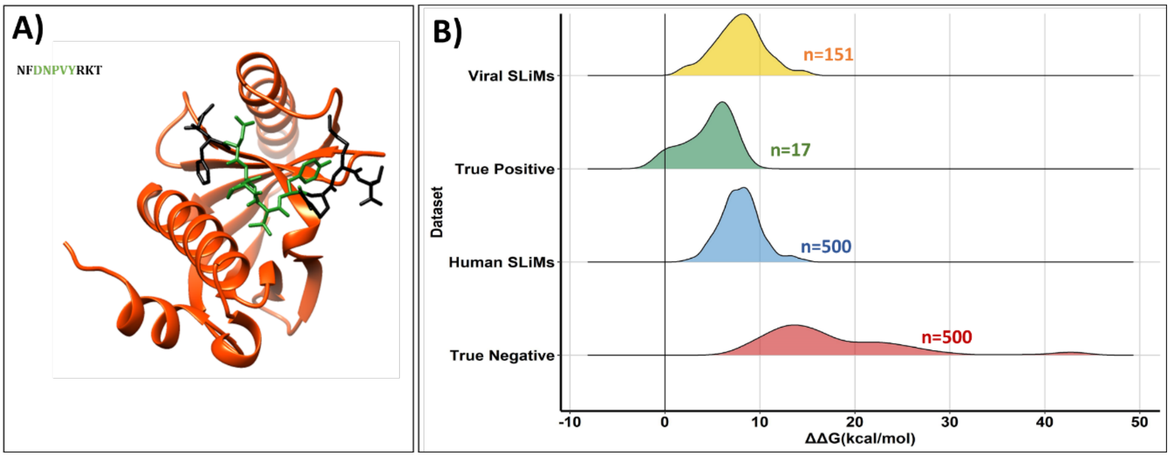


Figure 5.12. Peptide exchange experiment using known 3D DMI (PDB ID: 1NTV) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of LIG_PTB_APO_2 DMI complex (PDB ID: 1NTV), **B)** Binding energy differences (ΔΔG) of different peptide datasets bound to native LIG_PTB_APO_2 complex.

4- LIG_SH3_2

The LIG_SH3_2 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.13, Figure 5.2A).** True positive SLiMs showed less destabilisation than true negatives (Wilcoxon $p < 0.05$). All true positive SLiM occurrences were predicted to destabilise the structure (PDB ID: 1CKA, ΔΔG > 0) while 80 human SLiM occurrences were found to be stabilising (ΔΔG > 0). The true positives were found to be more destabilisation than random human occurrences (Wilcoxon $p < 0.05$) **(Figure 5.2B)**. A total of 71 viral SLiM were stabilising (ΔΔG < 0) **(Table 5.2, Figure 5.2D)** and were more stabilising than the human occurrences (Wilcoxon $p < 0.05$) **(Figure 5.2C)**. Among these binding SLiMs, 4 occurrences (i.e. PLPPPR, PPAPRR, PPLPAK, PVPPPR) were mimicry candidates.

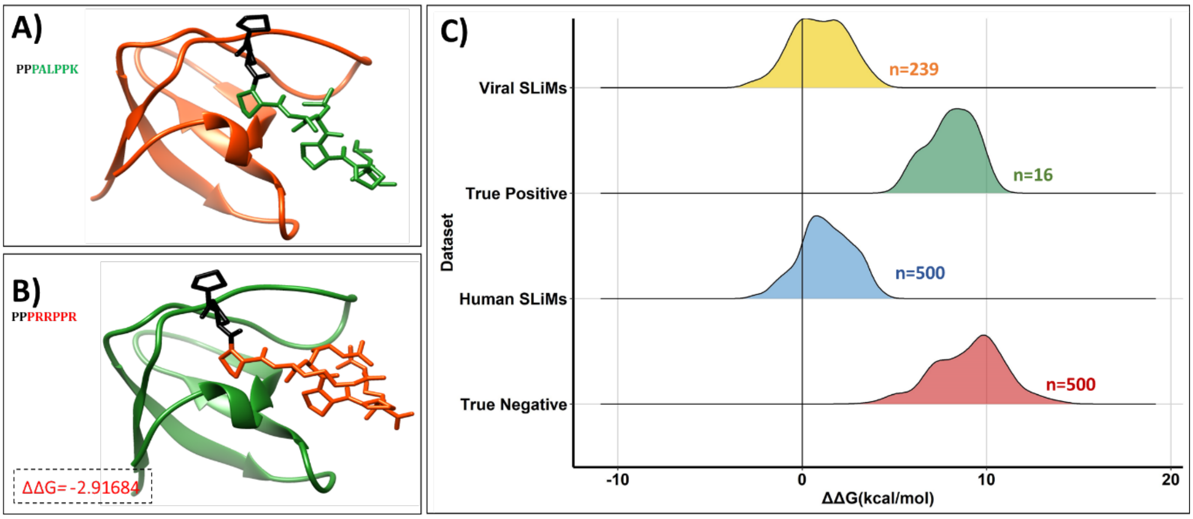


Figure 5.13. Peptide exchange experiment using known 3D DMI (PDB ID: 1CKA) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native LIG_SH3_2 complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native LIG_SH3_2 complex.

5- LIG_ULM_U2AF65_1

The LIG_ULM_U2AF65_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.14, Figure 5.2A).** True positive SLiMs showed more stabilisation than true negatives (Wilcoxon $p < 0.05$). A total of 30 human and only 1 true positive SLiM occurrence were predicted to stabilise the structure (PDB ID: 2PEH, ΔΔG < 0). The true positives were demonstrating higher destabilisation than random human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2B)**. Only 1 viral SLiM (RRRRWR) was found to be stabilising (ΔΔG < 0) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2C)**. None of the mimicry candidates were found to be stabilising.

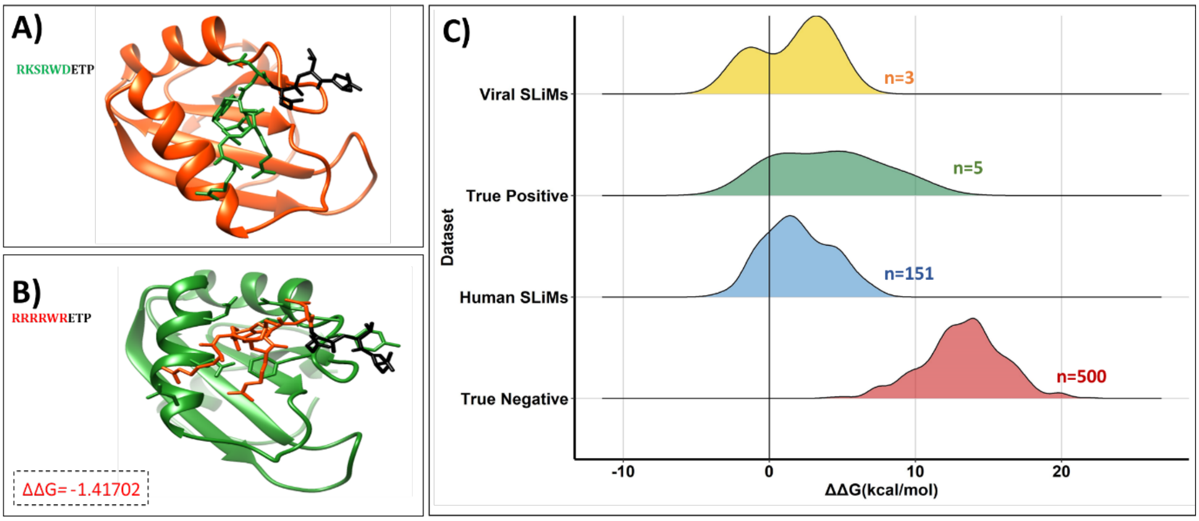


Figure 5.14. Peptide exchange experiment using known 3D DMI (PDB ID: 2PEH) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native LIG_ULM_U2AF65_1 complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native LIG_ULM_U2AF65_1 complex.

6- LIG_WD40_WDR5_WIN_1

LIG_WD40_WDR5_WIN_1 had 6 solved structures in PDB (i.e. PDB Id: 3UVM, 3UVN, 4CY1, 4ERZ, 4ES0 and 4ESG). The LIG_WD40_WDR5_WIN_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.16, Figure 5.2A).** True positive SLiMs showed more stabilisation than true negatives (Wilcoxon $p < 0.05$). Around 50% of the human and 1 true positive SLiM occurrence were stabilising (PDB ID: 3UVM, $\Delta\Delta G < 0$). The true positive were demonstrating higher destabilisation than the random human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2B)**. Only 1 viral SLiM was found to be stabilising (PDB ID: 3UVM, $\Delta\Delta G < 0$) **(Table 5.2, Figure 5.2D)** while none of viral SLiM showed stabilisation with other structures. These viral SLiMs could not be distinguished from human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2C)**.

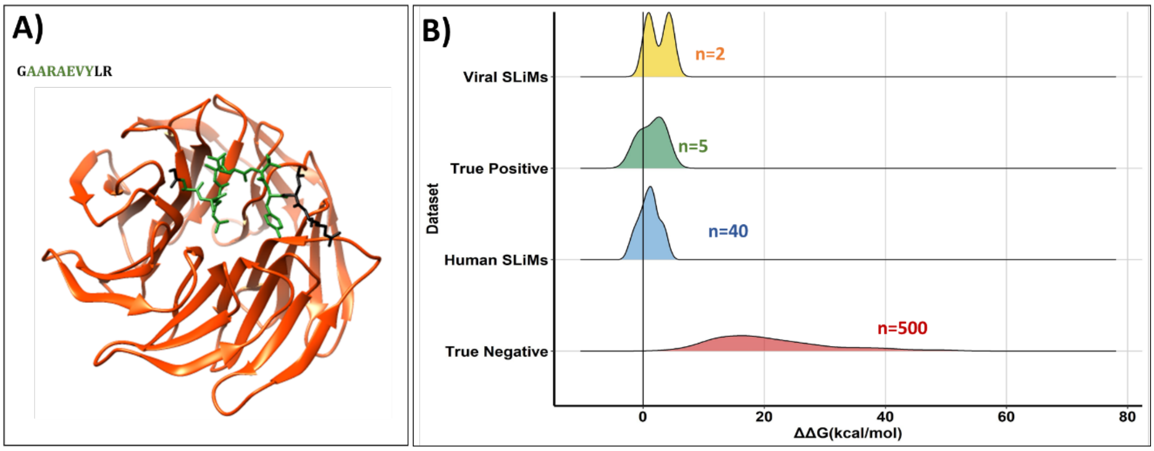


Figure 5.15. Peptide exchange experiment using known 3D DMI (PDB ID: 3UVM) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native LIG_WD40_WDR5_WIN_1 complex (3UVM), **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native LIG_WD40_WDR5_WIN_1 complex.

7- LIG_WW_1

The LIG_WW_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.16, Figure 5.2A).** True positive SLiMs showed less destabilisation than true negatives (Wilcoxon $p < 0.05$). All human and true positive SLiM occurrences were predicted to destabilise the structure (PDB ID: 1EG4, ΔΔG > 0), with the true positives demonstrating less stabilisation than random human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2B)**. Viral SLiM were also destabilising (ΔΔG > 0) and could not be distinguished from human occurrences (Wilcoxon $p > 0.05$) **(Figure 5.2C)**.

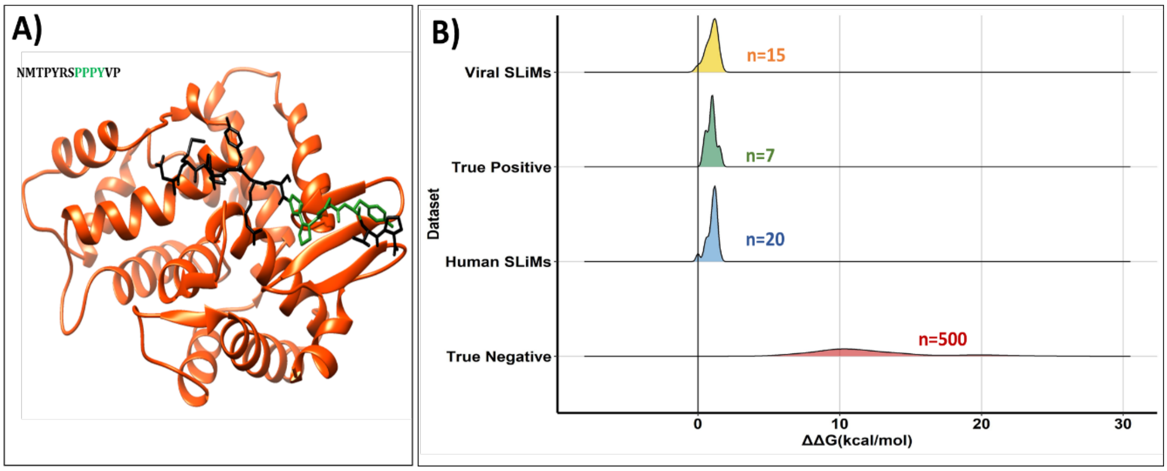


Figure 5.16. Peptide exchange experiment using known 3D DMI (PDB ID: 1EG4) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of LIG_WW_1 DMI complex (PDB ID: 1EG4), **B)** Binding energy differences (ΔΔG) of different peptide datasets bound to native LIG_WW_1 complex.

5.5.5 Targeting sites

Targeting site motifs are the sites within proteins that helps in recognition and binding of the proteins (Dinkel, Michael et al. 2012). 3D DMI complexes of two targeting ELMs (i.e. TRG_LYSEND_GGAACLL_1 and TRG_NLS_MONOEXTC_3) were selected to evaluate predicted viral instances of SLiMs.

1- TRG_LYSEND_GGAACLL_1

TRG_LYSEND_GGAACLL_1 had two solved structures in PDB (i.e. PDB ID: 1JUQ and 1JWG). The TRG_LYSEND_GGAACLL_1 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon $p < 0.05$) **(Figure 5.17, Figure 5.2A).** True positive SLiMs showed more stabilisation than true negatives (Wilcoxon $p < 0.05$). Around 50% of the true positive and human SLiM occurrences were predicted to stabilise the structure (PDB ID: 1JWG, $\Delta\Delta G > 0$). The true positives were demonstrating higher destabilisation than random human occurrences (Wilcoxon $p > 0.05$) (**Figure 5.2B)**. Both viral SLiMs: DRDLLD and DRNLLD were found to be stabilising ($\Delta\Delta G < 0$) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon $p > 0.05$)

(Figure 5.2C). On the other hand, all the viral SLiMs were destabilising (PDB ID: 1JUQ, ΔΔG > 0) and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**.

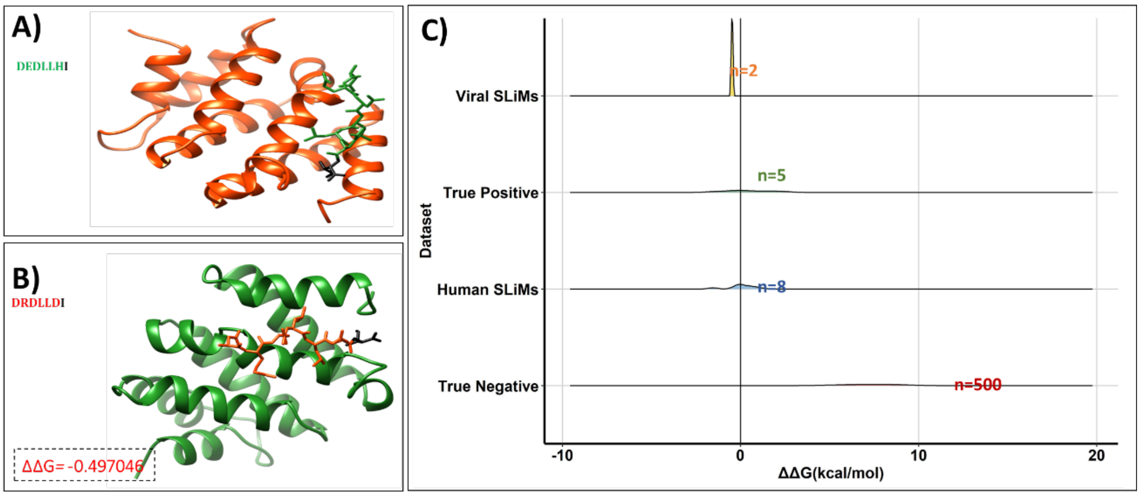


Figure 5.17. Peptide exchange experiment using known 3D DMI (PDB ID: 1JWG) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native TRG_LYSEND_GGAACLL_1 complex (PDB ID: 1JWG), **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native TRG_LYSEND_GGAACLL_1 complex.

2- TRG_NLS_MONOEXTC_3

The TRG_NLS_MONOEXTC_3 motif sequence contributed strongly to predicting binding affinity, with human occurrences of the motif showing smaller increases in binding energy than the random peptide true negative control (Wilcoxon p < 0.05) **(Figure 5.18, Figure 5.2A).** True positive SLiMs showed more stabilisation than true negatives (Wilcoxon p < 0.05). A total of 90 human and 2 true positive SLiM occurrences were predicted to stabilise the structure (PDB ID: 1EE4, ΔΔG < 0), with the true positives demonstrating higher destabilisation than random human occurrences (Wilcoxon p > 0.05) **(Figure 5.2B)**. A total of 17 viral SLiM were predicted to be stabilising (ΔΔG < 0) **(Table 5.2, Figure 5.2D)** and could not be distinguished from human occurrences (Wilcoxon p > 0.05) **(Figure 5.2C)**.

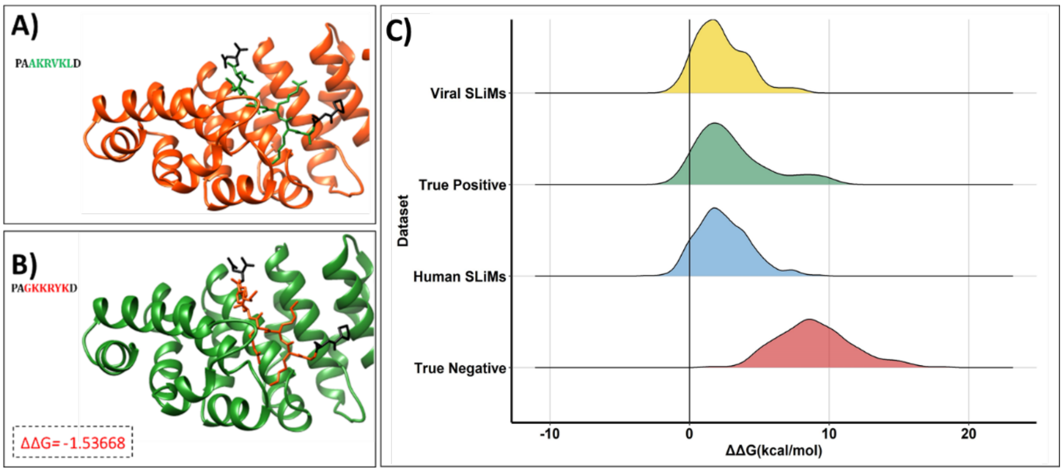


Figure 5.18. Peptide exchange experiment using known 3D DMI (PDB ID: 1EE4) complex.

X-axis shows the ΔΔG of the datasets, y-axis shows the frequency of the SLiMs. **A)** 3D structure of native TRG_NLS_MONOEXTC_3 complex, **B)** Top ranked predicted viral peptide bound to native complex, **C)** Binding energy differences (ΔΔG) of different datasets bound to native TRG_NLS_MONOEXTC_3 complex.

Table 5.2. Binding energy quantification of predicted viral SLiMs bound with native complexes.

ELM	PDB ID	Peptide	ΔΔG
DEG_SIAH_1	2FLU	VPTAEVRPG	-1.42749
		PPSAMVRPT	-0.837641
		RPRAAVAPC	-0.638656
		SPEARVPPG	-0.435736
		DPSAAVGPD	-0.0502077
DOC_ANK_TNKS_1	3TWW	KRNPPPGY	-2.60672
		PRRPPPGR	-2.4746
		PRPPPPGV	-2.21947
		RRHLPPGA	-2.11479
		PRGEPPGE	-1.29359
		LRRGPPGE	-1.26399
		TRPGPPGI*	-1.11788
		ARDPPPGA	-1.04761
		YRFAAPGE	-0.94632
		ARPGPPGI	-0.85382
		RRPLPDGT	-0.84212
		FRDRPDGV	-0.75661
		ARDYPDGE	-0.66953
		IRFVPDGS	-0.65491
		KRICPPGT	-0.45142
		LRSVPPGA	-0.23888
		YRFVAPGE	-0.1321
		FRSAPEGH	-0.05878
	3TWX	RRHLPPGA	-2.06377
		PRRPPPGR	-1.92777
		PRGEPPGE	-1.59202
		KRNPPPGY	-1.35326
		PRPPPPGV	-1.08908
		ARDPPPGA	-0.95831
		RRPLPDGT	-0.9175
		TRPGPPGI	-0.62013
		FRDRPDGV	-0.47364
		KRICPPGT	-0.37735
		LRRGPPGE	-0.29376
		IRFVPDGS	-0.25401
		ARDYPDGE	-0.23969
		ERQIPDGE	-0.18581
		KRNPPPGY	-0.0431
DOC_CYCLIN_1	1H24	RPLF	-0.806
		KFLM	-0.74952
		KPLV	-0.63684
		RFLP	-0.62226
		KPLI	-0.5936
		RPLI	-0.58142
		RPLM	-0.5802
		RPLP	-0.55168
		KPLY	-0.54005

		KLLI	-0.53769
		RPLV	-0.45703
		RFLL	-0.42992
		KPLP	-0.41306
		RYLP	-0.39823
		KPLM	-0.33716
		KYLP	-0.30821
		KLLY	-0.29493
		KYLL	-0.24751
		KYLF	-0.24149
		RMLY	-0.23371
		RQLP	-0.21754
		RLLY	-0.20771
		KRLP	-0.19793
		KRLY	-0.1799
		KVLF	-0.17337
		RMLL	-0.17087
		RKLP	-0.16837
		KFLL	-0.16428
		KKLP	-0.16402
		RHLP	-0.15951
		KRLF	-0.14818
		RFLV	-0.14378
		RRLP	-0.12972
		RLLM	-0.11178
		RVLP	-0.10935
		KVLI	-0.09815
		RGLP	-0.07398
		RILL	-0.06607
		KRLL	-0.06511
		KQLI	-0.0651
		RRLF	-0.04871
		KILV	-0.0403
		KMLL	-0.03455
		KPLI	-0.0338
		KFLV	-0.03228
		KVLL	-0.02996
		KHLP	-0.02929
		KKLI	-0.02336
		RRLL	-0.01526
		RRLI	-0.00904
		KLLY	-0.00281
		RRLY	-0.0027
		RLLP	-5.61E-05
LIG_PTAP_UEV_1	3OBQ	RPTAPP	-2.8675
		IPTAPP	-2.84484
		LPTAPP	-2.78543
		QPTAPP	-2.66342
		RPSAPP	-2.55268

		XPSAPA	-2.34341
		RPTAPS	-2.31539
		VPTAPP	-2.21693
		EPSAPP	-2.20307
		VPSAPP	-2.11169
		SPTAPP	-2.10624
		RPTAPF	-2.00816
		APTAPP	-1.8349
		RPTAPF	-1.82629
		RPTAPL	-1.54596
		WPSAPE	-1.52059
		LPSAPE	-1.39835
		YPTAPA	-1.28903
		RPTAPS	-1.14059
		TPTAPL	-1.0665
		RPTAPT	-0.86517
		KPTAPT	-0.71092
		EPTAPQ	-0.65948
		APSAPM	-0.60975
		RPSAPA	-0.60142
		APTAPL	-0.60111
		EPTAPS	-0.56065
		LPSAPT	-0.53908
		TPSAPS	-0.43513
		VPTAPA	-0.29788
		TPSAPT	-0.26448
		SPTAPS	-0.12407
		VPSAPG	-0.01472
LIG_SH3_2	1CKA	PRRPPR	-2.91684
		PPLPSR	-2.80718
		PKLPPR	-2.58222
		PPLPPR	-2.53579
		PPKPPR	-2.41982
		PPRPKR	-1.97147
		PERPPR	-1.9081
		PKKPPR	-1.86624
		PQLPPR	-1.81122
		PPPPPR	-1.71979
		PPLPYR	-1.60075
		PPQPPR	-1.58239
		PPPPAR	-1.50299
		PPLPPK	-1.30033
		PKPPPR	-1.22129
		PPPPRR	-1.20963
		PARPTR	-1.19562
		PNKPHR	-1.17332
		PPPPKR	-1.15073
		PPRPTR	-1.11246
		PHRPTR	-1.08734

		PPAPPR	-1.04964
		PPPPTR	-0.98882
		PKIPKR	-0.97325
		PPAPKR	-0.9608
		PRPPRR	-0.94436
		PRLPAR	-0.94277
		PPPPSR	-0.91557
		PPPPQR	-0.78104
		PRVPRR	-0.74394
		PMRPLR	-0.66238
		PPPPGR	-0.6112
		PGPPPR	-0.60723
		PDLPGR	-0.59703
		PQKPPR	-0.59513
		PPVPYR	-0.59207
		PLIPYR	-0.58539
		PPGPRR	-0.56151
		PLPPPR*	-0.55425
		PRPPSR	-0.54541
		PLKPTR	-0.51366
		PPAPAR	-0.49732
		PVKPRR	-0.49166
		PPAPRR*	-0.43815
		PNPPGR	-0.39125
		PPLPAK*	-0.38319
		PGRPTR	-0.36409
		PRKPLR	-0.35501
		PLRPSR	-0.33971
		PPGPPR	-0.3391
		PIPPPR	-0.28179
		PPSPPR	-0.27983
		PRRPRR	-0.25823
		PEAPPR	-0.25004
		PEQPSR	-0.24212
		PPPPER	-0.22663
		PAQPPR	-0.20711
		PPVPIR	-0.19738
		PRTPSR	-0.17877
		PPTPQR	-0.12391
		PRVPGR	-0.11668
		PPAPSR	-0.09931
		PPPPPK	-0.09345
		PVTPKK	-0.08891
		PAVPSR	-0.04916
		PVPPPR*	-0.04384
		PRTPAR	-0.04063
		PRQPAR	-0.02158
		PQKPPR	-0.02031
		PGKPSR	-0.01699

		PLEPPR	-0.00806
TRG_LYSEND_GGAACLL_1	1JWG	DRDLLD	-0.497046
		DRNLLD	-0.446507
TRG_NLS_MONOEXTC_3	1EE4	GKKRYK	-1.53668
		RKRRKR	-0.99513
		RKKLKR	-0.9379
		QKRPRR	-0.792
		PKKRLR	-0.7424
		PKKVKR	-0.73008
		PKKKRK	-0.67413
		KKKRKR	-0.29727
		RKKPRK	-0.25185
		RKKRQR	-0.20796
		TRKRIR	-0.19874
		IKKRFK	-0.17045
		LKKLKK	-0.15317
		GKKRKR	-0.1072
		RKKLKP	-0.08233
		MKRFRK	-0.0621
		TKKKYK	-0.06186

*mimicry candidates from predicted DMI dataset.

5.6 Discussion

The analysis in this chapter is a pilot study where known DMI knowledge was combined with predicted SLiMs/DMI knowledge to discover new SLiM occurrences through peptide exchange experiment. More specifically, to see if binding energy quantification of predicted SLiM occurrences with the native DMI complexes could be used to discriminate binding vs non-binding motifs.

5.6.1 Binding energy quantification does not necessarily help in validating DMIs.

Once the structural data was prepared, it was run through FoldX with peptide datasets to identify binding vs non-binding SLiMs. Four major types of ELMs (i.e. DOC, DEG, LIG and TRG) were evaluated for this pilot study to find new SLiM occurrences through swapping the already bound peptide with predicted peptides. Most of the viral SLiM occurrences destabilised the native structures. The binding energy quantification was assessed based on the ΔΔG values. A ΔΔG > 0 suggests that the binding peptide is destabilising the structure and a ΔΔG <0 shows that the peptide is stabilising the structure (Schymkowitz, Borg et al. 2005). The predicted viral SLiMs were found to be stabilising for 10 ELM classes based on their ΔΔG (i.e. DEG_SIAH_1, DOC_ANK_TNKS_1, DOC_CYCLIN_1, LIG_PTAP_UEV_1, LIG_PTB_APO_2, LIG_SH3_2, LIG_ULM_U2AF65_1, LIG_WW_1, TRG_LYSEND_GGAACLL_1 and TRG_NLS_MONOEXTC_3). Out of 23 analysed structures, over 50% demonstrated stabilisation with the predicted viral peptides (ΔΔG < 0). A Wilcoxon rank test was carried out to see if the designed approach was working in principle. Four different comparisons were done: 1) random human occurrences vs true negative peptides which was done to see if FoldX could detect motif specificity, 2) true positive SLiMs vs random human occurrences to see if TP shows stronger binding than random human SLiMs and, 3) viral vs random human occurrences to see if viral SLiMs were stronger binders than the random human SLiMs, 4) TP vs TN to see specificity of SLiMs **(Figure 5.2).** For the potential validation first two conditions needed to be significant and for the actual validation of the predictions, all four conditions needed to be significant. However, if third condition wasn't found significant, it still suggests that individual viral SLiM predictions can still be good candidates.

5.6.1.1 Can FoldX detect general motif specificity

The first question was to see whether FoldX detect general motif specificity (motifs versus random peptides). It was expected that all the true negative peptides will destabilise the complex. This was indeed the case as all the true negative peptides showed very high ΔΔG (ΔΔG >0), indicating that none of them were able to bind with the native structure through maintaining its stability. This was expected as SLiM interactions are highly specific (SLiM binds with a specific interaction partner) (Zarrinpar, Park et al. 2003). The statistical significant difference of the human vs true negative peptides **(Figure 5.2)** demonstrated that true negatives were more destabilising than the random human occurrences which in turn shows that FoldX can detect general motif specificity.

5.6.1.2 Do TP generally show stronger binding than random human SLiMs?

It was expected that all true positive occurrences would be stronger binders than the random human SLiMs and would be more stabilising than the human SLiMs. Not all true positive SLiMs were found to be stabilising the native complex. For example, none of the true positive occurrence showed stabilisation (ΔΔG < 0) with the LIG_WW_1 (PDB ID: 1EG4), DEG_KELCH_KEAP1_1 (PDB ID: 2FLU), DOC_AGCK_PIF_1 (PDB ID: 1O6L) and LIG_LIR_GEN_1 (PDB ID: 3DOW). An apparent lack of significant difference between TP and random human occurrences was observed for most of the complexes **(Figure 5.2)** suggesting that not all true positives can be good binders. The reason could be the mechanisms involved in binding (i.e. localisation and/or timing of expression) which prevent binding *in vivo*. For some motifs (i.e. DEG_KELCH_KEAP1_1 (2FLU), DEG_SIAH_1 (2A25), DOC_ANK_TNKS_1 (3TWU), DOC_CYCLIN_1 (1H24), LIG_LIR_GEN_1 (3DOW), LIG_PTB_APO_2 (1NTV), LIG_SH3_2 (1CKA)), a significant difference (P-value < 0.05) between TP and random human occurrences was observed showing that the approach was working in principle **(Figure 5.2)** and TPs were stronger binders than the random human occurrences. This raises the possibility of using the approach for *de-novo* SLiM predictions. For example, the predicted TPs can be compared with the proteome background to find a new motif class. In general, it can be said that stabilisation alone as a parameter cannot be used for validation of predictions and better tools needs to be utilized to investigate binding energies.

5.6.1.3 Are viral SLiMs stronger binders than the random human SLiMs

Given that the condition (TP vs Human) was true sometimes, the next question was to see if viral SLiMs stronger binders were than the random human SLiMs. This was not the case as an apparent lack of significant difference between viral and human SLiMs was observed for almost all complexes except for LIG_SH3_2 complex. This in turn supports that viral predictions for LIG_SH3_2 could be enriched for TPs. The lack of significant difference between viral and human also suggests that there might be false positives in the viral SLiMs. In general, most of the viral SLiMs showed ΔΔG >0 suggesting that they were destabilising the native complex. As SLiM predictions come with high false positive rates, it was likely that many of the predicted SLiMs would be false predictions.

On a general note, it can be said that this approach is not universally useful in its current form, but still helped in identifying some individual candidates of interest. The binding energy changes of identified SLiMs were quite effective (ΔΔG < 0) suggesting that these could possibly be ideal candidates for future validations through experimental techniques.

5.6.1.4 Binding energy changes of different ELM classes

First, degron motifs (i.e. DEG_KELCH_KEAP1_1 and DEG_SIAH_1) were analysed. Degron motifs are SLiMs which are embedded in modular proteins and are used by E3 ubiquitin ligases to target proteins for degradation. These motifs are known to mediate several cellular functions including monitoring cellular hypoxia and progression through cell cycle. In general, these motifs are responsible for prevention of protein dysfunction through eliminating proteins which are no longer required (Meszaros, Kumar et al. 2017). In case of DEG_KELCH_KEAP1_1, none of the known occurrences showed ΔΔG < 0 indicating that all these SLiMs were destabilising the native structure. The reason could be that the available known occurrences are not specific to that domain/structure. This was also the case with the random human occurrences as all of them destabilised the structure (ΔΔG > 0). Similarly, all the viral SLiMs also destabilized the native structure. On the other hand, around 50% known as well as predicted viral SLiMs were stabilising the DEG_SIAH_1 complex (ΔΔG < 0) **(Figure 5.4).**

In case of docking motifs, firstly, binding energy changes were evaluated for DOC_AGCK_PIF_1 complex. None of the known SLiMs showed stabilisation while few random human SLiMs showed stabilisation (ΔΔG <0). All the viral SLiMs destabilised the native complex indicating that those might be just random occurrences **(Figure 5.5).** On the other hand, both human and viral SLiMs showed effective binding with the DOC_CYCLIN_1 complex. Around ¼ of the predicted viral SLiMs demonstrated stabilisation (ΔΔG < 0) with the DOC_CYCLIN_1 **(Figure 5.9, Table 5.2)** complex. Similarly, effectively binding SLiMs were identified for DOC_ANK_TNKS_1 complex. DOC_ANK_TNKS_1 had three solved (i.e. 3TWU, 3TWW and 3TWX) DMI complexes in PDB. It was observed that not all SLiMs were binding in same manner with these three complexes. Some tend to bind more effectively with the 3TWU **(Figure 5.6, Table 5.2**Error! Reference source not found.**),** some with 3TWW **(Figure 5.7, Table 5.2)** and some with the 3TWX **(Figure 5.8, Table 5.2)**. The overall binding energy changes (ΔΔG) were different for all these complexes. One possible explanation could be that, efficiency of a SLiM binding to the globular domain might be dependent on different factors (e.g. quality of the solved structure, method used to solve structure, stability of the complex etc.).

Looking at the ligand motifs (i.e. LIG_LIR_GEN_1, LIG_PTAP_UEV_1, LIG_PTB_APO_2, LIG_SH3_2, LIG_ULM_U2AF65_1, LIG_WD40_WDR5_WIN_1 and LIG_WW_1), did not identify any effectively binding viral SLiMs for the LIG_LIR_GEN_1 **(Figure 5.10)**, LIG_WW_1 **(Figure 5.11)** and LIG_PTB_APO_2 **(Figure 5.12),** but found many stabilising SLiMs for LIG_PTAP_UEV_1 **(Figure 5.13, Table 5.2),** LIG_SH3_2 **(Figure 5.14, Table 5.2),** LIG_WD40_WDR5_WIN_1 **(Figure 5.15, Table 5.2)** and LIG_ULM_U2AF65_1 **(Figure 5.16, Table 5.2)**. All true positive occurrences were destabilising the native complex of LIG_SH3_2 while few human random occurrences demonstrated stabilisation of the complex. The reason of true positive destabilising the structure could be motif sub-specificity for different SH3 domains and thus they might have weaker-than-random binding. LIG_SH3_2 showed significant difference (P-value < 0.05) for all the three conditions. This was the only motif class which showed significant difference for all three conditions **(Figure 5.2)** indicating that the predicted viral SLiMs could be potential candidates for validation of the DMIs. The interaction between different LIG_SH3_2 motifs and SH3 domains has previously been reported where high specificity of the SH3 interactions was shown (Zarrinpar, Park et al. 2003). This analysis also showed this specificity where LIG_SH3_2 motifs interacted effectively with the native SH3 complex.

Many stabilising SLiMs were identified for the targeting motifs (i.e. TRG_LYSEND_GGAACLL_1 **(Figure 5.17, Table 5.2)** and TRG_NLS_MONOEXTC_3 **(Figure 5.18, Table 5.2)**. Most of the known and random human occurrences showed ΔΔG below 0 indicating that these SLiMs were binding through maintaining the stability of the complex.

5.6.2 Limitations/Issues of the approach

In-silico peptide exchange experiment wasn't as successful as expected to initially validate the predictions. The reason could be the several issues faced during the analysis. One prominent limitation of this approach was the loss of data for the analysis. A big chunk of structural data was lost because of the limitation of FoldX dealing with non-standard amino acids. Most of the native structures had peptides with some non-standard amino acids (e.g. phosphorylated residues). Since it was a pilot study therefore, structures which were suitable for FoldX and did not had any such issues were analysed. Unfortunately, a big proportion of data (~76%) was lost during this step. This issue could be fixed through using other computational tools which are more suitable for dealing with non-standard residues.

Another possible solution would be to try to remove phosphorylation from the bound peptides so that they can be used with FoldX. Another big drawback of this analysis was destabilisation of the native complex. Most of the predictions destabilised the native complex (ΔΔG > 0). Comparison with the controls clearly shows that slightly destabilising energies still indicate good binding.

The pipeline designed for this analysis did not work as effective as it was expected. The reason of this pipeline not working does not necessarily mean that FoldX is not suitable for such analysis rather it might have been influenced by high false discovery rate of the predictions. A better and improved pipeline can help in improving such analysis for example, using molecular docking approach through Autodock (Morris, Huey et al. 2009). Other tools which also enable peptide exchange are FlexPepDock (Raveh, London et al. 2011) where high -resolution models of peptide-protein complexes between flexible peptides and proteins are generated and Rosetta (Das and Baker 2008; Alford, Leaver-Fay et al. 2017) where peptide-protein binding can be done. The typical docking approach would be to remove already bound peptide and replace it with the predicted peptides by selecting the binding pocket in protein. This kind of approach would eventually help in finding the binding mode of the individual predictions through giving atomic level information. Such type of analysis would require high computational resources and time.

The final step would be to do *in-vivo* and *in-vitro* experimental validations. Different generic experiments (i.e. mutation analysis, alanine scanning, co-immunoprecipitation, pull-down, two hybrid and colocalization) can be used to validate motifs (Gibson, Dinkel et al. 2015). Studies have been conducted in past to validate motifs through phage display experiments for example, in one study co-immunoprecipitation in conjunction with isothermal titration calorimetry was used to validate motifs. (i.e. interactions between DxxLL motifs and GGA1 VHS domains) (Davey, Seo et al. 2017). One possible way to validate the identified SLiMs is through a mutagenesis experiment where the already bound peptide can be mutated through replacing with the identified SLiMs. Moreover, for the ELM classes where all the three conditions were found to be significant, protein can be expressed with the viral SLiMs that were found to be stronger than the random human occurrences.

5.6.3 Conclusion

The work done in this chapter was a pilot study to see if *in-silico* peptide exchange experiment could help in discriminating binding vs non-binding SLiMs and to see if this could be used for the initial validation of the DMIs. The designed pipeline/approach was not found to be efficient in terms of initial validation of the predictions but helped in identifying some effectively binding SLiMs. As the predictions come with high false discovery rate, it was expected that not many mimicry candidates would bind with the known complexes. This was indeed the case as most of the mimicry candidates didn't show stabilisation with the native complex and showed ΔΔG above 0. This emphasizes the need that better approach/pipeline should be adapted to reduce false discovery rate. Despite the inefficiency of designed pipeline, it helped in screening SLiMs which stabilised the native complex. These SLiMs predictions could be real and need to be validated through wet lab experiments where the actual binding affinity of the peptides could be evaluated.

6 Chapter 6: Conclusions

The main aim of my thesis was to study molecular mimicry in viruses through short linear motifs (SLiMs). Viruses mimic SLiMs in host proteins and establish transient interactions known as domain-motif interactions (DMIs). The current number of known DMIs in databases like ELM (Dinkel, Van Roey et al. 2016) and 3did (Mosca, Ceol et al. 2014) is underrepresented therefore, the first objective of my thesis was to design a pipeline that could predict DMIs from the protein-protein interactions (PPIs) data.

6.1 SLiMEnrich

The idea was to combine the known/predicted motif, domain composition with the PPI data to predict DMIs and to evaluate enrichment. The designed pipeline was converted into an interactive online application known as SLiMEnrich (Idrees, Perez-Bercoff et al. 2018). SLiMEnrich mainly helps in evaluating whether a PPI source is enriched in terms of capturing DMIs and helps in predicting new DMIs. SLiMEnrich works in three possible ways: 1) ELMi-Protein strategy where a PPI source is mapped to known DMI data from ELM to see how well it is capturing DMIs, 2) ELMc-Protein strategy where the motif containing protein is linked to its known interactor protein (domain containing protein) via ELM class, 3) ELMc-Domain strategy where the motif containing protein is linked to ELM, ELM to Pfam domain and Pfam domain to its corresponding proteins. SLiMEnrich by default works with the known SLiMs in the ELM databases but it can essentially work with any types of motif predictions for instance, SLiMProb predictions can be used to predict new DMIs. The usage of SLiMEnrich is not limited to DMI prediction rather it can be used for other purposes as well. For example, it can be used to predict other types of interactions (i.e. domain-domain interactions (DDIs), Protein-motif interactions (PMIs)) provided that the files are in right format. It can be used to find the optimal stringency of the interactions and it can also be used by systems biologists to see if their proteins/PPI of interest are enriched in a given PPI data. A command-line version of SLiMEnrich is also available which can deal with large datasets with easy to handle commands.

6.2 PPI datasets as a source of DMI and DDI predictions

After developing the SLiMEnrich, my next objective was to use it to evaluate how well different publicly available PPI data sources were capturing different sorts of interactions (i.e. DMIs and DDIs). SLiMEnrich's main algorithm works by calculating enrichment of DMIs in PPI datasets through comparing it with a random set of PPIs. The reason of calculating enrichment based on random pair of PPIs than the whole interactome was to assure whether a dataset can be a good source of DMIs/DDIs. Once assured that a dataset is capturing DMIs/DDIs, this information can be used to design further downstream analysis for example, calculating enrichment by comparing with the whole network and finding how enriched is a DMI/DDI given an interactome.

For this purpose, 10 publicly available human interactomes were compared to see how well they were capturing DMIs and DDIs and which high-throughput method (i.e. Y2H, AP-MS and CoFrac-MS) was better at capturing these interactions. It was seen that all datasets were significantly enriched in terms of capturing DMIs and DDIs. Both Y2H and AP-MS looked promising in terms of capturing these interactions while CoFrac-MS wasn't as compared to them. The one thing that would make such type of analysis would be availability of a comprehensive database based on high-throughput interactions. In future, this analysis can be extended through including more PPI datasets and also through comparing different low-throughput methods to see how they capture different sorts of interactions.

6.3 Predicting motif mimicry in viruses

Once it was assured the designed pipeline was working and human interactome was capturing DMIs, my next objective was to utilize SLiMEnrich to study molecular mimicry in viruses. For this purpose, two viral-human PPI datasets (i.e. PHISTO and VirHostNet2.0) were evaluated to see if they were capturing DMIs. Both datasets captured really low number of known DMIs. This wasn't surprising as the number of known virus human DMIs are low in ELM. This in general emphasizes the need to discover more DMIs and incorporate them in databases like ELM. It should also be noted that ELM is not currently documenting all the known viral SLiMs and more viral data needs to be incorporated in ELM. Just like human interactome analysis, both Y2H and AP-MS showed significant enrichment in terms of capturing DMIs. The comparison of different viral subtypes leads to prediction of new DMIs, but it should be noted that these predictions comes with high false discovery rate. This was the reason that I didn’t investigate individual results. This sort of analysis can become more powerful and reliable through including more PPI data for different viral subtypes and through adding additional filtrations to reduce FDR rate. Finally, I discovered new SLiMs through the integration of human and viral-human interactomes. For this analysis, I integrated HI-II-14 and PHISTO datasets. In future, more comprehensive PPI datasets can be used to improve this analysis and discover more SLiMs.

6.4 Structural validation of predictions

Once the DMI prediction analysis was done, my next objective was to develop a pipeline for the initial validation of the predicted SLiMs/DMIs. I designed a pilot study where I used *in-silico* peptide exchange experiment through exchanging the already bound SLiM in the known DMI complex with the predicted SLiMs. On a general note, this approach wasn’t as efficient as expected. Several issues were faced during this analysis i.e. I lost large proportion of the 3D data as most of them had some non-standard amino-acid residues in their bound peptide sequence. It can be said that this analysis could be improved by using other tools i.e. Rosetta or some molecular docking and dynamics techniques to ensure the binding mode of the predictions. Even though, the designed pipeline didn’t work as expected, it helped in screening few high confidence peptides which needs validation through computational as well as experimental techniques.

6.5 Future goals

On a final note, I would like to say that the analysis done can be improved through designing better pipelines that could help in reducing the FDR rate. I have future aims to improve SLiMEnrich through adding more filtration steps i.e. to remove post-translational modifications from the analysis, to focus on certain types of motif types and to remove highly abundant domains/motifs from the analysis etc. I would also like to do some experimental validation of the high confidence SLiMs I gained from the peptide exchange experiments. In future, I plan on extending my PhD work for more in-depth analysis of the individual DMIs.

7 References

Accardi, R., R. Rubino, M. Scalise, T. Gheit, N. Shahzad, M. Thomas, L. Banks, C. Indiveri, B. S. Sylla, R. A. Cardone, S. J. Reshkin and M. Tommasino (2011). "E6 and E7 from human papillomavirus type 16 cooperate to target the PDZ protein Na/H exchange regulatory factor 1." J Virol **85**(16): 8208-8216.

Akiva, E., G. Friedlander, Z. Itzhaki and H. Margalit (2012). "A dynamic view of domain-motif interactions." PLoS Comput Biol **8**(1): e1002341.

Alanis-Lobato, G., M. A. Andrade-Navarro and M. H. Schaefer (2017). "HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks." Nucleic Acids Res **45**(D1): D408-D414.

Alford, R. F., A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme and J. J. Gray (2017). "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design." J Chem Theory Comput **13**(6): 3031-3048.

Aloy, P. and R. B. Russell (2006). "Structural systems biology: modelling protein interactions." Nat Rev Mol Cell Biol **7**(3): 188-197.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi and L. S. Yeh (2004). "UniProt: the Universal Protein knowledgebase." Nucleic Acids Res **32**(Database issue): D115-119.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bailey, T. L. and M. Gribskov (1997). "Score distributions for simultaneous matching to multiple motifs." J Comput Biol **4**(1): 45-59.

Bairoch, A. (1993). "The PROSITE dictionary of sites and patterns in proteins, its current status." Nucleic Acids Res **21**(13): 3097-3103.

Balla, S., V. Thapar, S. Verma, T. Luong, T. Faghri, C. H. Huang, S. Rajasekaran, J. J. del Campo, J. H. Shinn, W. A. Mohler, M. W. Maciejewski, M. R. Gryk, B. Piccirillo, S. R. Schiller and M. R. Schiller (2006). "Minimotif Miner: a tool for investigating protein function." Nat Methods **3**(3): 175-177.

Barnes, B., M. Karimloo, A. Schoenrock, D. Burnside, E. Cassol, A. Wong, F. Dehne, A. Golshani and J. R. Green (2016). "Predicting Novel Protein-Protein Interactions Between the HIV-1 Virus and Homo Sapiens." 2016 IEEE EMBS International Student Conference (ISC).

Baspinar, A., E. Cukuroglu, R. Nussinov, O. Keskin and A. Gursoy (2014). "PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes." Nucleic Acids Res **42**(Web Server issue): W285-289.

Becerra, A., V. A. Bucheli and P. A. Moreno (2017). "Prediction of virus-host protein-protein interactions mediated by short linear motifs." BMC Bioinformatics **18**(1): 163.

Benedict, C. A., P. S. Norris and C. F. Ware (2002). "To kill or be killed: viral evasion of apoptosis." Nat Immunol **3**(11): 1013-1018.

Berlow, R. B., H. J. Dyson and P. E. Wright (2018). "Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation." J Mol Biol **430**(16): 2309-2320.

Betel, D., K. E. Breitkreuz, R. Isserlin, D. Dewar-Darch, M. Tyers and C. W. Hogue (2007). "Structure-templated predictions of novel protein interactions from sequence information." PLoS Comput Biol **3**(9): 1783-1789.

Bhattacharyya, R. P., A. Remenyi, B. J. Yeh and W. A. Lim (2006). "Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits." Annu Rev Biochem **75**: 655-680.

Bhowmick, P., M. Guharoy and P. Tompa (2015). "Bioinformatics Approaches for Predicting Disordered Protein Motifs." Adv Exp Med Biol **870**: 291-318.

Bjorkholm, P. and E. L. Sonnhammer (2009). "Comparative analysis and unification of domain-domain interaction networks." Bioinformatics **25**(22): 3020-3025.

Blikstad, C. and Y. Ivarsson (2015). "High-throughput methods for identification of protein-protein interactions involving short linear motifs." Cell Commun Signal **13**: 38.

Boyen, P., D. Van Dyck, F. Neven, R. C. van Ham and A. D. van Dijk (2011). "SLIDER: a generic metaheuristic for the discovery of correlated motifs in protein-protein interaction networks." IEEE/ACM Trans Comput Biol Bioinform **8**(5): 1344-1357.

Brannetti, B. and M. Helmer-Citterich (2003). "iSPOT: A web tool to infer the interaction specificity of families of protein modules." Nucleic Acids Res **31**(13): 3709-3711.

Breuer, K., A. K. Foroushani, M. R. Laird, C. Chen, A. Sribnaia, R. Lo, G. L. Winsor, R. E. Hancock, F. S. Brinkman and D. J. Lynn (2013). "InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation." Nucleic Acids Res **41**(Database issue): D1228-1233.

Brown, K. R. and I. Jurisica (2005). "Online predicted human interaction database." Bioinformatics **21**(9): 2076-2082.

Bruckner, A., C. Polge, N. Lentze, D. Auerbach and U. Schlattner (2009). "Yeast two-hybrid, a powerful tool for systems biology." Int J Mol Sci **10**(6): 2763-2788.

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

Ceol, A., A. Chatr-aryamontri, E. Santonico, R. Sacco, L. Castagnoli and G. Cesareni (2007). "DOMINO: a database of domain-peptide interactions." Nucleic Acids Res **35**(Database issue): D557-560.

Chaurushiya, M. S., C. E. Lilley, A. Aslanian, J. Meisenhelder, D. C. Scott, S. Landry, S. Ticau, C. Boutell, J. R. Yates, 3rd, B. A. Schulman, T. Hunter and M. D. Weitzman (2012). "Viral E3 ubiquitin ligase-mediated degradation of a cellular E3: viral mimicry of a cellular phosphorylation mark targets the RNF8 FHA domain." Mol Cell **46**(1): 79-90.

Chemes, L. B., G. de Prat-Gay and I. E. Sanchez (2015). "Convergent evolution and mimicry of protein linear motifs in host-pathogen interactions." Curr Opin Struct Biol **32**: 91-101.

Chemes, L. B., I. E. Sanchez and G. de Prat-Gay (2011). "Kinetic recognition of the retinoblastoma tumor suppressor by a specific protein target." J Mol Biol **412**(2): 267-284.

Chou, M. F. and D. Schwartz (2011). "Biological sequence motif discovery using motif-x." Curr Protoc Bioinformatics **Chapter 13**: Unit 13 15-24.

Corbi-Verge, C. and P. M. Kim (2016). "Motif mediated protein-protein interactions as drug targets." Cell Commun Signal **14**: 8.

Cowley, M. J., M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi and J. Wu (2012). "PINA v2.0: mining interactome modules." Nucleic Acids Res **40**(Database issue): D862-865.

D'haeseleer, P. (2006). "How does DNA sequence motif discovery work?" Nature Biotechnology **24**(8): 959-961.

D'haeseleer, P. (2006). "What are DNA sequence motifs?" Nature Biotechnology **24**(4): 423-425.

Damian, R. T. (1964). "Molecular mimicry: antigen sharing by parasite and host and its consequences." The American Naturalist **98**(900): 129-149.

Das, J. and H. Yu (2012). "HINT: High-quality protein interactomes and their applications in understanding human disease." BMC Syst Biol **6**: 92.

Das, R. and D. Baker (2008). "Macromolecular modeling with rosetta." Annu Rev Biochem **77**: 363-382.

Davey, N. E., J. L. Cowan, D. C. Shields, T. J. Gibson, M. J. Coldwell and R. J. Edwards (2012). "SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions." Nucleic Acids Res **40**(21): 10628-10641.

Davey, N. E., N. J. Haslam, D. C. Shields and R. J. Edwards (2011). "SLiMSearch 2.0: biological context for short linear motifs in proteins." Nucleic Acids Res **39**(Web Server issue): W56-60.

Davey, N. E., M. H. Seo, V. K. Yadav, J. Jeon, S. Nim, I. Krystkowiak, C. Blikstad, D. Dong, N. Markova, P. M. Kim and Y. Ivarsson (2017). "Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome." FEBS J **284**(3): 485-498.

Davey, N. E., D. C. Shields and R. J. Edwards (2006). "SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent." Nucleic Acids Res **34**(12): 3546-3554.

Davey, N. E., D. C. Shields and R. J. Edwards (2009). "Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery." Bioinformatics **25**(4): 443-450.

Davey, N. E., G. Trave and T. J. Gibson (2011). "How viruses hijack cell regulation." Trends in Biochemical Sciences **36**(3): 159-169.

Davey, N. E., K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel and T. J. Gibson (2012). "Attributes of short linear motifs." Mol Biosyst **8**(1): 268-281.

de Castro, E., C. J. A. Sigrist, A. Gattiker, V. Bulliard, P. S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch and N. Hulo (2006). "ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins." Nucleic Acids Research **34**: W362-W365.

de Chassey, B., L. Meyniel-Schicklin, J. Vonderscher, P. Andre and V. Lotteau (2014). "Virus-host interactomics: new insights and opportunities for antiviral drug discovery." Genome Med **6**(11): 115.

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

De Las Rivas, J. and C. Fontanillo (2010). "Protein-protein interactions essentials: key concepts to building and analyzing interactome networks." PLoS Comput Biol **6**(6): e1000807.

De Las Rivas, J. and C. Fontanillo (2012). "Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell." Brief Funct Genomics **11**(6): 489-496.

Deane, C. M., L. Salwinski, I. Xenarios and D. Eisenberg (2002). "Protein interactions: two methods for assessment of the reliability of high throughput observations." Mol Cell Proteomics **1**(5): 349-356.

Diella, F., N. Haslam, C. Chica, A. Budd, S. Michael, N. P. Brown, G. Trave and T. J. Gibson (2008). "Understanding eukaryotic linear motifs and their role in cell signaling and regulation." Front Biosci **13**: 6580-6603.

Dinh, H., S. Rajasekaran and J. Davila (2012). "qPMS7: a fast algorithm for finding (l, d)-motifs in DNA and protein sequences." PLoS One **7**(7): e41425.

Dinkel, H., S. Michael, R. J. Weatheritt, N. E. Davey, K. Van Roey, B. Altenberg, G. Toedt, B. Uyar, M. Seiler, A. Budd, L. Jodicke, M. A. Dammert, C. Schroeter, M. Hammer, T. Schmidt, P. Jehl, C. McGuigan, M. Dymecka, C. Chica, K. Luck, A. Via, A. Chatr-Aryamontri, N. Haslam,

G. Grebnev, R. J. Edwards, M. O. Steinmetz, H. Meiselbach, F. Diella and T. J. Gibson (2012). "ELM--the database of eukaryotic linear motifs." Nucleic Acids Res **40**(Database issue): D242-251.

Dinkel, H. and H. Sticht (2007). "A computational strategy for the prediction of functional linear peptide motifs in proteins." Bioinformatics **23**(24): 3297-3303.

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

Dinkel, H., K. Van Roey, S. Michael, M. Kumar, B. Uyar, B. Altenberg, V. Milchevskaya, M. Schneider, H. Kuhn, A. Behrendt, S. L. Dahl, V. Damerell, S. Diebel, S. Kalman, S. Klein, A. C. Knudsen, C. Mader, S. Merrill, A. Staudt, V. Thiel, L. Welti, N. E. Davey, F. Diella and T. J. Gibson (2016). "ELM 2016-data update and new functionality of the eukaryotic linear motif resource." Nucleic Acids Res **44**(D1): D294-300.

Disfani, F. M., W. L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, A. K. Dunker, V. N. Uversky and L. Kurgan (2012). "MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins." Bioinformatics **28**(12): i75-83.

Dogruel, M., T. A. Down and T. J. Hubbard (2008). "NestedMICA as an ab initio protein motif discovery tool." BMC Bioinformatics **9**: 19.

Dosztanyi, Z., V. Csizmok, P. Tompa and I. Simon (2005). "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content." Bioinformatics **21**(16): 3433-3434.

Dosztanyi, Z., B. Meszaros and I. Simon (2009). "ANCHOR: web server for predicting protein binding regions in disordered proteins." Bioinformatics **25**(20): 2745-2746.

Durmus, S. and K. O. Ulgen (2017). "Comparative interactomics for virus-human protein-protein interactions: DNA viruses versus RNA viruses." FEBS Open Bio **7**(1): 96-107.

Durmus Tekir, S., T. Cakir, E. Ardic, A. S. Sayilirbas, G. Konuk, M. Konuk, H. Sariyer, A. Ugurlu, I. Karadeniz, A. Ozgur, F. E. Sevilgen and K. O. Ulgen (2013). "PHISTO: pathogen-host interaction search tool." Bioinformatics **29**(10): 1357-1358.

Durmus Tekir, S., T. Cakir and K. O. Ulgen (2012). "Infection Strategies of Bacterial and Viral Pathogens through Pathogen-Human Protein-Protein Interactions." Front Microbiol **3**: 46.

Duro, N., M. Miskei and M. Fuxreiter (2015). "Fuzziness endows viral motif-mimicry." Mol Biosyst **11**(10): 2821-2829.

Dyer, M. D., T. M. Murali and B. W. Sobral (2007). "Computational prediction of host-pathogen protein-protein interactions." Bioinformatics **23**(13): i159-166.

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

Edwards, R. J. (2013). "SLiMSuite software package." Retrieved 20/01/2016, 2016, from http://www.southampton.ac.uk/~re1u06/software/packages/slimsuite/.

Edwards, R. J., N. E. Davey, K. O'Brien and D. C. Shields (2012). "Interactome-wide prediction of short, disordered protein interaction motifs in humans." Mol Biosyst **8**(1): 282-295.

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

Edwards, R. J., N. E. Davey and D. C. Shields (2008). "CompariMotif: quick and easy comparisons of sequence motifs." Bioinformatics **24**(10): 1307-1309.

Edwards, R. J. and N. Palopoli (2015). "Computational prediction of short linear motifs from protein sequences." Methods Mol Biol **1268**: 89-141.

Emamjomeh, A., B. Goliaei, J. Zahiri and R. Ebrahimpour (2014). "Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method." Molecular Biosystems **10**(12): 3147-3154.

Encinar, J. A., G. Fernandez-Ballester, I. E. Sanchez, E. Hurtado-Gomez, F. Stricher, P. Beltrao and L. Serrano (2009). "ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs." Bioinformatics **25**(18): 2418-2424.

Evans, P., W. Dampier, L. Ungar and A. Tozeren (2009). "Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs." Bmc Medical Genomics **2**.

Fang, C., T. Noguchi, D. Tominaga and H. Yamana (2013). "MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation." BMC Bioinformatics **14**: 300.

Finlay, B. B. and G. McFadden (2006). "Anti-immunology: evasion of the host immune system by bacterial and viral pathogens." Cell **124**(4): 767-782.

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

Finn, R. D., B. L. Miller, J. Clements and A. Bateman (2014). "iPfam: a database of protein family and domain interactions found in the Protein Data Bank." Nucleic Acids Res **42**(Database issue): D364-373.

Formstecher, E., S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J. A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M. P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulo, F. Perez, A. Plessis, C. Rosse, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis and L. Daviet (2005). "Protein interaction mapping: a Drosophila case study." Genome Res **15**(3): 376-384.

Franzosa, E. A. and Y. Xia (2011). "Structural principles within the human-virus protein-protein interaction network." Proceedings of the National Academy of Sciences of the United States of America **108**(26): 10538-10543.

Freund, C., R. Kuhne, H. Yang, S. Park, E. L. Reinherz and G. Wagner (2002). "Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules." EMBO J **21**(22): 5985-5995.

Frith, M. C., N. F. Saunders, B. Kobe and T. L. Bailey (2008). "Discovering sequence motifs with arbitrary insertions and deletions." PLoS Comput Biol **4**(4): e1000071.

Ganesh, R., D. A. Siegele and T. R. Ioerger (2003). "MOPAC: motif finding by preprocessing and agglomerative clustering from microarrays." Pac Symp Biocomput: 41-52.

Ganti, K., J. Broniarczyk, W. Manoubi, P. Massimi, S. Mittal, D. Pim, A. Szalmas, J. Thatte, M. Thomas, V. Tomaic and L. Banks (2015). "The Human Papillomavirus E6 PDZ Binding Motif: From Life Cycle to Malignancy." Viruses **7**(7): 3530-3551.

Garamszegi, S., E. A. Franzosa and Y. Xia (2013). "Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks." PLoS Pathog **9**(12): e1003778.

Gibson, T. J. (2009). "Cell regulation: determined to signal discrete cooperation." Trends Biochem Sci **34**(10): 471-482.

Gibson, T. J., H. Dinkel, K. Van Roey and F. Diella (2015). "Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad." Cell Commun Signal **13**: 42.

Goll, J., S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb and P. Uetz (2008). "MPIDB: the microbial protein interaction database." Bioinformatics **24**(15): 1743-1744.

Gouw, M., S. Michael, H. Samano-Sanchez, M. Kumar, A. Zeke, B. Lang, B. Bely, L. B. Chemes, N. E. Davey, Z. Deng, F. Diella, C. M. Gurth, A. K. Huber, S. Kleinsorg, L. S. Schlegel, N. Palopoli, K. V. Roey, B. Altenberg, A. Remenyi, H. Dinkel and T. J. Gibson (2017). "The eukaryotic linear motif resource - 2018 update." Nucleic Acids Res.

Grant, C. E., T. L. Bailey and W. S. Noble (2011). "FIMO: scanning for occurrences of a given motif." Bioinformatics **27**(7): 1017-1018.

Green, T. J., R. Cox, J. Tsao, M. Rowse, S. H. Qiu and M. Luo (2014). "Common Mechanism for RNA Encapsidation by Negative-Strand RNA Viruses." Journal of Virology **88**(7): 3766-3775.

Guirimand, T., S. Delmotte and V. Navratil (2015). "VirHostNet 2.0: surfing on the web of virus/host molecular interactions data." Nucleic Acids Res **43**(Database issue): D583-587.

Hagai, T., A. Azia, A. Toth-Petroczy and Y. Levy (2011). "Intrinsic disorder in ubiquitination substrates." J Mol Biol **412**(3): 319-324.

Halehalli, R. R. and H. A. Nagarajaram (2015). "Molecular principles of human virus protein-protein interactions." Bioinformatics **31**(7): 1025-1033.

Harak, C. and V. Lohmann (2015). "Ultrastructure of the replication sites of positive-strand RNA viruses." Virology **479**: 418-433.

Hauser, R., S. Blasche, T. Dokland, E. Haggard-Ljungquist, A. von Brunn, M. Salas, S. Casjens, I. Molineux and P. Uetz (2012). "Bacteriophage protein-protein interactions." Adv Virus Res **83**: 219-298.

Havugimana, P. C., G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V. U. Dar, A. Bezginov,

G. W. Clark, G. C. Wu, S. J. Wodak, E. R. Tillier, A. Paccanaro, E. M. Marcotte and A. Emili (2012). "A census of human soluble protein complexes." Cell **150**(5): 1068-1081.

Hecker, C. M., M. Rabiller, K. Haglund, P. Bayer and I. Dikic (2006). "Specification of SUMO1- and SUMO2-interacting motifs." J Biol Chem **281**(23): 16117-16127.

Hein, M. Y., N. C. Hubner, I. Poser, J. Cox, N. Nagaraj, Y. Toyoda, I. A. Gak, I. Weisswange, J. Mansfeld, F. Buchholz, A. A. Hyman and M. Mann (2015). "A human interactome in three quantitative dimensions organized by stoichiometries and abundances." Cell **163**(3): 712-723.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman and R. Apweiler (2004). "IntAct: an open source molecular interaction database." Nucleic Acids Res **32**(Database issue): D452-455.

Hon, L. S. and A. N. Jain (2006). "A deterministic motif finding algorithm with application to the human genome." Bioinformatics **22**(9): 1047-1054.

Hornbeck, P. V., J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham and M. Sullivan (2012). "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse." Nucleic Acids Res **40**(Database issue): D261-270.

Hu, S., E. Song, R. Tian, S. Ma, T. Yang, Y. Mu, Y. Li, C. Shao, S. Gao and Y. Gao (2009). "Systematic analysis of a simple adaptor protein PDZK1: ligand identification, interaction and functional prediction of complex." Cell Physiol Biochem **24**(3-4): 231-242.

Huang, H., P. B. McGarvey, B. E. Suzek, R. Mazumder, J. Zhang, Y. Chen and C. H. Wu (2011). "A comprehensive protein-centric ID mapping service for molecular data integration." Bioinformatics **27**(8): 1190-1191.

Hulo, N., A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni and C. J. Sigrist (2006). "The PROSITE database." Nucleic Acids Res **34**(Database issue): D227-230.

Hung, A. Y. and M. Sheng (2002). "PDZ domains: structural modules for protein complex assembly." J Biol Chem **277**(8): 5699-5702.

Huttlin, E. L., R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P. Gygi and J. W. Harper (2017). "Architecture of the human interactome defines protein communities and disease networks." Nature.

Idrees, S., A. Perez-Bercoff and R. J. Edwards (2018). "Correction: SLiMEnrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions." PeerJ **6**.

Iyer, L. M., L. Aravind and E. V. Koonin (2001). "Common origin of four diverse families of large eukaryotic DNA viruses." J Virol **75**(23): 11720-11734.

Jean Beltran, P. M., J. D. Federspiel, X. Sheng and I. M. Cristea (2017). "Proteomics and integrative omic approaches for understanding host-pathogen interactions and infectious diseases." Mol Syst Biol **13**(3): 922.

Jonassen, I., J. F. Collins and D. G. Higgins (1995). "Finding flexible patterns in unaligned protein sequences." Protein Sci **4**(8): 1587-1595.

Kalathur, R. K. R., J. P. Pinto, M. A. Hernandez-Prieto, R. S. R. Machado, D. Almeida, G. Chaurasia and M. E. Futschik (2014). "UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks." Nucleic Acids Research **42**(D1): D408-D414.

Kamburov, A., U. Stelzl, H. Lehrach and R. Herwig (2013). "The ConsensusPathDB interaction database: 2013 update." Nucleic Acids Res **41**(Database issue): D793-800.

Kaneko, T., L. Li and S. S. Li (2008). "The SH3 domain--a family of versatile peptide- and protein-recognition module." Front Biosci **13**: 4938-4952.

Kazlauskas, D., M. Krupovic and C. Venclovas (2016). "The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes." Nucleic Acids Res **44**(10): 4551-4564.

Kazlauskas, D. and C. Venclovas (2011). "Computational analysis of DNA replicases in double-stranded DNA viruses: relationship with the genome size." Nucleic Acids Res **39**(19): 8291-8305.

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

Kiel, C. and L. Serrano (2014). "Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations." Mol Syst Biol **10**: 727.

Kim, E. D., A. Sabharwal, A. R. Vetta and M. Blanchette (2010). "Predicting direct protein interactions from affinity purification mass spectrometry data." Algorithms Mol Biol **5**: 34.

Kim, M. S., S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda and A. Pandey (2014). "A draft map of the human proteome." Nature **509**(7502): 575-581.

Kim, Y., B. Min and G. S. Yi (2012). "IDDI: integrated domain-domain interaction and protein interaction analysis system." Proteome Sci **10 Suppl 1**: S9.

Kinoshita, K. and H. Nakamura (2005). "Identification of the ligand binding sites on the molecular surface of proteins." Protein Sci **14**(3): 711-718.

Kohm, A. P., K. G. Fuller and S. D. Miller (2003). "Mimicking the way to autoimmunity: an evolving theory of sequence and structural homology." Trends Microbiol **11**(3): 101-105.

Konecna, A., R. Frischknecht, J. Kinter, A. Ludwig, M. Steuble, V. Meskenaite, M. Indermuhle, M. Engel, C. Cen, J. M. Mateos, P. Streit and P. Sonderegger (2006). "Calsyntenin-1 docks vesicular cargo to kinesin-1." Mol Biol Cell **17**(8): 3651-3663.

Koonin, E. V., M. Krupovic and N. Yutin (2015). "Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses." DNA Habitats and Their Rna Inhabitants **1341**: 10-24.

Krupovic, M. and P. Forterre (2015). "Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes." Ann N Y Acad Sci **1341**: 41-53.

Krupovic, M. and P. Forterre (2015). "Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes." DNA Habitats and Their Rna Inhabitants **1341**: 41-53.

Krystkowiak, I. and N. E. Davey (2017). "SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions." Nucleic Acids Res.

Kuchaiev, O., M. Rasajski, D. J. Higham and N. Przulj (2009). "Geometric de-noising of protein-protein interaction networks." PLoS Comput Biol **5**(8): e1000454.

Lam, H. Y., P. M. Kim, J. Mok, R. Tonikian, S. S. Sidhu, B. E. Turk, M. Snyder and M. B. Gerstein (2010). "MOTIPS: automated motif analysis for predicting targets of modular protein domains." BMC Bioinformatics **11**: 243.

Langeberg, L. K. and J. D. Scott (2015). "Signalling scaffolds and local organization of cellular behaviour." Nat Rev Mol Cell Biol **16**(4): 232-244.

Lee, J. O., A. A. Russo and N. P. Pavletich (1998). "Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7." Nature **391**(6670): 859-865.

Lee, T., A. N. Hoofnagle, Y. Kabuyama, J. Stroud, X. Min, E. J. Goldsmith, L. Chen, K. A. Resing and N. G. Ahn (2004). "Docking motif interactions in MAP kinases revealed by hydrogen exchange mass spectrometry." Mol Cell **14**(1): 43-55.

Leung, H. C. and F. Y. Chin (2006). "Finding motifs from all sequences with and without binding sites." Bioinformatics **22**(18): 2217-2223.

Leung, H. C., M. H. Siu, S. M. Yiu, F. Y. Chin and K. W. Sung (2009). "Clustering-based approach for predicting motif pairs from protein interaction data." J Bioinform Comput Biol **7**(4): 701-716.

Li, D., T. Wei, C. M. Abbott and D. Harrich (2013). "The unexpected roles of eukaryotic translation elongation factors in RNA virus replication and pathogenesis." Microbiol Mol Biol Rev **77**(2): 253-266.

Li, X., M. Wu, C. K. Kwoh and S. K. Ng (2010). "Computational approaches for detecting protein complexes from protein interaction networks: a survey." BMC Genomics **11 Suppl 1**: S3.

Licata, L., L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli and G. Cesareni (2012). "MINT, the molecular interaction database: 2012 update." Nucleic Acids Res **40**(Database issue): D857-861.

Lieber, D. S., O. Elemento and S. Tavazoie (2010). "Large-scale discovery and characterization of protein regulatory motifs in eukaryotes." PLoS One **5**(12): e14444.

Lim, J., T. Hao, C. Shaw, A. J. Patel, G. Szabo, J. F. Rual, C. J. Fisk, N. Li, A. Smolyar, D. E. Hill, A. L. Barabasi, M. Vidal and H. Y. Zoghbi (2006). "A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration." Cell **125**(4): 801-814.

Linding, R., L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. Jorgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe and T. Pawson (2007). "Systematic discovery of in vivo phosphorylation networks." Cell **129**(7): 1415-1426.

Liu, X. and R. Marmorstein (2007). "Structure of the retinoblastoma protein bound to adenovirus E1A reveals the molecular basis for viral oncoprotein inactivation of a tumor suppressor." Genes Dev **21**(21): 2711-2716.

Liu, Y., E. Wimmer and A. V. Paul (2009). "Cis-acting RNA elements in human and animal plus-strand RNA viruses." Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms **1789**(9-10): 495-517.

Lodish, L. H., A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnel (2000). Viruses: Structure, Function, and Uses. New York, W. H. Freeman.

Lubovac, Z., J. Gamalielsson and B. Olsson (2006). "Combining functional and topological properties to identify core modules in protein interaction networks." Proteins **64**(4): 948-959.

Luck, K., G. M. Sheynkman, I. Zhang and M. Vidal (2017). "Proteome-Scale Human Interactomics." Trends Biochem Sci **42**(5): 342-354.

Lum, K. K. and I. M. Cristea (2016). "Proteomic approaches to uncovering virus-host protein interactions during the progression of viral infection." Expert Rev Proteomics.

Lyon, K. F., X. Cai, R. J. Young, A. A. Mamun, S. Rajasekaran and M. R. Schiller (2018). "Minimotif Miner 4: a million peptide minimotifs and counting." Nucleic Acids Res **46**(D1): D465-D470.

Maere, S., K. Heymans and M. Kuiper (2005). "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks." Bioinformatics **21**(16): 3448-3449.

Martin, A. C. (2005). "Mapping PDB chains to UniProtKB entries." Bioinformatics **21**(23): 4297-4301.

McDowall, M. D., M. S. Scott and G. J. Barton (2009). "PIPs: human protein-protein interaction prediction database." Nucleic Acids Res **37**(Database issue): D651-656.

Mertens, P. (2004). "The dsRNA viruses." Virus Res **101**(1): 3-13.

Meszaros, B., M. Kumar, T. J. Gibson, B. Uyar and Z. Dosztanyi (2017). "Degrons in cancer." Sci Signal **10**(470).

Mi, T., J. C. Merlin, S. Deverasetty, M. R. Gryk, T. J. Bill, A. W. Brooks, L. Y. Lee, V. Rathnayake, C. A. Ross, D. P. Sargeant, C. L. Strong, P. Watts, S. Rajasekaran and M. R. Schiller (2012). "Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences." Nucleic Acids Res **40**(Database issue): D252-260.

Miller, M. L., L. J. Jensen, F. Diella, C. Jorgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, G. Cesareni, T. Pawson, B. E. Turk, M. B. Yaffe, S. Brunak and R. Linding (2008). "Linear motif atlas for phosphorylation-dependent signaling." Sci Signal **1**(35): ra2.

Mooney, C., G. Pollastri, D. C. Shields and N. J. Haslam (2012). "Prediction of short linear protein binding regions." J Mol Biol **415**(1): 193-204.

Morris, G. M., R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson (2009). "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility." J Comput Chem **30**(16): 2785-2791.

Mosca, R., A. Ceol, A. Stein, R. Olivella and P. Aloy (2014). "3did: a catalog of domain-based interactions of known three-dimensional structure." Nucleic Acids Res **42**(Database issue): D374-379.

Moya, A., S. F. Elena, A. Bracho, R. Miralles and E. Barrio (2000). "The evolution of RNA viruses: A population genetics view." Proc Natl Acad Sci U S A **97**(13): 6967-6973.

Neduva, V. and R. B. Russell (2005). "Linear motifs: Evolutionary interaction switches." Febs Letters **579**(15): 3342-3345.

Neduva, V. and R. B. Russell (2006). "DILIMOT: discovery of linear motifs in proteins." Nucleic Acids Res **34**(Web Server issue): W350-355.

Neduva, V. and R. B. Russell (2006). "Peptides mediating interaction networks: new leads at last." Curr Opin Biotechnol **17**(5): 465-471.

Ng, T. F., R. Marine, C. Wang, P. Simmonds, B. Kapusinszky, L. Bodhidatta, B. S. Oderinde, K. E. Wommack and E. Delwart (2012). "High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage." J Virol **86**(22): 12161-12175.

Nguyen Ba, A. N., B. J. Yeh, D. van Dyk, A. R. Davidson, B. J. Andrews, E. L. Weiss and A. M. Moses (2012). "Proteome-wide discovery of evolutionary conserved sequences in disordered regions." Sci Signal **5**(215): rs1.

Nourry, C., S. G. Grant and J. P. Borg (2003). "PDZ domain proteins: plug and play!" Sci STKE **2003**(179): RE7.

Nygren, P. J. and J. D. Scott (2015). "Therapeutic strategies for anchored kinases and phosphatases: exploiting short linear motifs and intrinsic disorder." Front Pharmacol **6**: 158.

O'Brien, K. T., N. J. Haslam and D. C. Shields (2013). "SLiMScape: a protein short linear motif analysis plugin for Cytoscape." BMC Bioinformatics **14**: 224.

Obenauer, J. C., L. C. Cantley and M. B. Yaffe (2003). "Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs." Nucleic Acids Res **31**(13): 3635-3641.

Oldstone, M. B. (1998). "Molecular mimicry and immune-mediated diseases." FASEB J **12**(13): 1255-1265.

Orchard, S., M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni and H. Hermjakob (2014). "The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases." Nucleic Acids Res **42**(Database issue): D358-363.

Ortin, J. and J. Martin-Benito (2015). "The RNA synthesis machinery of negative-stranded RNA viruses." Virology **479**: 532-544.

Oughtred, R., A. Chatr-Aryamontri, B. J. Breitkreutz, C. S. Chang, J. M. Rust, C. L. Theesfeld, S. Heinicke, A. Breitkreutz, D. Chen, J. Hirschman, N. Kolas, M. S. Livstone, J. Nixon, L. O'Donnell, L. Ramage, A. Winter, T. Reguly, A. Sellam, C. Stark, L. Boucher, K. Dolinski and M. Tyers (2016). "BioGRID: A Resource for Studying Biological Interactions in Yeast." Cold Spring Harb Protoc **2016**(1): pdb top080754.

Palopoli, N., K. T. Lythgow and R. J. Edwards (2015). "QSLiMFinder: improved short linear motif prediction using specific query protein data." Bioinformatics **31**(14): 2284-2293.

Pancsa, R. and M. Fuxreiter (2012). "Interactions via intrinsically disordered regions: what kind of motifs?" IUBMB Life **64**(6): 513-520.

Pawson, T. and R. Linding (2005). "Synthetic modular systems--reverse engineering of signal transduction." FEBS Lett **579**(8): 1808-1814.

Pawson, T., M. Raina and P. Nash (2002). "Interaction domains: from simple binding events to complex cellular behavior." FEBS Lett **513**(1): 2-10.

Peng, X., J. Wang, W. Peng, F. X. Wu and Y. Pan (2017). "Protein-protein interactions: detection, reliability assessment and applications." Brief Bioinform **18**(5): 798-819.

Pichlmair, A., K. Kandasamy, G. Alvisi, O. Mulhern, R. Sacco, M. Habjan, M. Binder, A. Stefanovic, C. A. Eberle, A. Goncalves, T. Burckstummer, A. C. Muller, A. Fauster, C. Holze, K. Lindsten, S. Goodbourn, G. Kochs, F. Weber, R. Bartenschlager, A. G. Bowie, K. L. Bennett, J. Colinge and G. Superti-Furga (2012). "Viral immune modulators perturb the human molecular network by common and unique strategies." Nature **487**(7408): 486-490.

Pitre, S., M. Alamgir, J. R. Green, M. Dumontier, F. Dehne and A. Golshani (2008). "Computational methods for predicting protein-protein interactions." Adv Biochem Eng Biotechnol **110**: 247-267.

Plewczynski, D., S. Basu and I. Saha (2012). "AMS 4.0: consensus prediction of post-translational modifications in protein sequences." Amino Acids **43**(2): 573-582.

Poltronieri, P., B. Sun and M. Mallardo (2015). "RNA Viruses: RNA Roles in Pathogenesis, Coreplication and Viral Load." Curr Genomics **16**(5): 327-335.

Prakash, A., M. Blanchette, S. Sinha and M. Tompa (2004). "Motif discovery in heterogeneous sequence data." Pac Symp Biocomput: 348-359.

Prieto, C. and J. De Las Rivas (2006). "APID: Agile Protein Interaction DataAnalyzer." Nucleic Acids Res **34**(Web Server issue): W298-302.

Prytuliak, R., M. Volkmer, M. Meier and B. H. Habermann (2017). "HH-MOTiF: de novo detection of short linear motifs in proteins by Hidden Markov Model comparisons." Nucleic Acids Res **45**(W1): W470-W477.

Raghavachari, B., A. Tasneem, T. M. Przytycka and R. Jothi (2008). "DOMINE: a database of protein domain interactions." Nucleic Acids Res **36**(Database issue): D656-661.

Rajagopala, S. V., P. Sikorski, A. Kumar, R. Mosca, J. Vlasblom, R. Arnold, J. Franca-Koh, S. B. Pakala, S. Phanse, A. Ceol, R. Hauser, G. Siszler, S. Wuchty, A. Emili, M. Babu, P. Aloy, R. Pieper and P. Uetz (2014). "The binary protein-protein interaction landscape of Escherichia coli." Nat Biotechnol **32**(3): 285-290.

Rao, V. B. and M. Feiss (2015). "Mechanisms of DNA Packaging by Large Double-Stranded DNA Viruses." Annu Rev Virol **2**(1): 351-378.

Raveh, B., N. London, L. Zimmerman and O. Schueler-Furman (2011). "Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors." PLoS One **6**(4): e18934.

Rigaut, G., A. Shevchenko, B. Rutz, M. Wilm, M. Mann and B. Seraphin (1999). "A generic protein purification method for protein complex characterization and proteome exploration." Nat Biotechnol **17**(10): 1030-1032.

Rigoutsos, I. and A. Floratos (1998). "Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm." Bioinformatics **14**(1): 55-67.

Rolland, T., M. Tasan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A. R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J. C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A. L. Barabasi, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth and M. Vidal (2014). "A proteome-scale map of the human interactome network." Cell **159**(5): 1212-1226.

Rosario, K., K. A. Mettel, B. E. Benner, R. Johnson, C. Scott, S. Z. Yusseff-Vanegas, C. C. M. Baker, D. L. Cassill, C. Storer, A. Varsani and M. Breitbart (2018). "Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates." PeerJ **6**: e5761.

Rose, P. W., A. Prlic, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y. P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo,

H. Yang, J. Y. Young, C. Zardecki, H. M. Berman and S. K. Burley (2017). "The RCSB protein data bank: integrative view of protein, gene and 3D structural information." Nucleic Acids Res **45**(D1): D271-D281.

Rozenblatt-Rosen, O., R. C. Deo, M. Padi, G. Adelmant, M. A. Calderwood, T. Rolland, M. Grace, A. Dricot, M. Askenazi, M. Tavares, S. J. Pevzner, F. Abderazzaq, D. Byrdsong, A. R. Carvunis, A. A. Chen, J. Cheng, M. Correll, M. Duarte, C. Fan, M. C. Feltkamp, S. B. Ficarro, R. Franchi, B. K. Garg, N. Gulbahce, T. Hao, A. M. Holthaus, R. James, A. Korkhin, L. Litovchick, J. C. Mar, T. R. Pak, S. Rabello, R. Rubio, Y. Shen, S. Singh, J. M. Spangle, M. Tasan, S. Wanamaker, J. T. Webber, J. Roecklein-Canfield, E. Johannsen, A. L. Barabasi, R. Beroukhim, E. Kieff, M. E. Cusick, D. E. Hill, K. Munger, J. A. Marto, J. Quackenbush, F. P. Roth, J. A. DeCaprio and M. Vidal (2012). "Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins." Nature **487**(7408): 491-495.

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

Sarkar, D., T. Jana and S. Saha (2015). "LMPID: a manually curated database of linear motifs mediating protein-protein interactions." Database (Oxford) **2015**.

Schlundt, A., J. Sticht, K. Piotukh, D. Kosslick, N. Jahnke, S. Keller, M. Schuemann, E. Krause and C. Freund (2009). "Proline-rich sequence recognition: II. Proteomics analysis of Tsg101 ubiquitin-E2-like variant (UEV) interactions." Mol Cell Proteomics **8**(11): 2474-2486.

Schuster-Bockler, B. and A. Bateman (2007). "Reuse of structural domain-domain interactions in protein networks." BMC Bioinformatics **8**: 259.

Schymkowitz, J., J. Borg, F. Stricher, R. Nys, F. Rousseau and L. Serrano (2005). "The FoldX web server: an online force field." Nucleic Acids Res **33**(Web Server issue): W382-388.

Seet, B. T., I. Dikic, M. M. Zhou and T. Pawson (2006). "Reading protein modifications with interaction domains." Nat Rev Mol Cell Biol **7**(7): 473-483.

Segura-Cabrera, A., C. A. Garcia-Perez, X. Guo and M. A. Rodriguez-Perez (2013). "A viral-human interactome based on structural motif-domain interactions captures the human infectome." PLoS One **8**(8): e71526.

Segura-Cabrera, A., C. A. Garcia-Perez, X. W. Guo and M. A. Rodriguez-Perez (2013). "A Viral-Human Interactome Based on Structural Motif-Domain Interactions Captures the Human Infectome." Plos One **8**(8).

Seiler, M., A. Mehrle, A. Poustka and S. Wiemann (2006). "The 3of5 web application for complex and comprehensive pattern matching in protein sequences." BMC Bioinformatics **7**: 144.

Seo, M. H. and P. M. Kim (2018). "The present and the future of motif-mediated protein-protein interactions." Curr Opin Struct Biol **50**: 162-170.

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

Shelton, H. and M. Harris (2008). "Hepatitis C virus NS5A protein binds the SH3 domain of the Fyn tyrosine kinase with high affinity: mutagenic analysis of residues within the SH3 domain that contribute to the interaction." Virol J **5**: 24.

Siddharthan, R., E. D. Siggia and E. van Nimwegen (2005). "PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny." PLoS Comput Biol **1**(7): e67.

Simonis, N., J. F. Rual, A. R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab, S. Cevik, N. Klitgord, C. Fan, P. Braun, N. Li, N. Ayivi-Guedehoussou, E. Dann, N. Bertin, D. Szeto, A. Dricot, M. A. Yildirim, C. Lin, A. S. de Smet, H. L. Kao, C. Simon, A. Smolyar, J. S. Ahn, M. Tewari, M. Boxem, S. Milstein, H. Yu, M. Dreze, J. Vandenhaute, K. C. Gunsalus, M. E. Cusick, D. E. Hill, J. Tavernier, F. P. Roth and M. Vidal (2009). "Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network." Nat Methods **6**(1): 47-54.

Sinha, S. (2003). "Discriminative motifs." J Comput Biol **10**(3-4): 599-615.

Sinha, S., M. Blanchette and M. Tompa (2004). "PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences." BMC Bioinformatics **5**: 170.

Sinha, S. and M. Tompa (2000). "A statistical method for finding transcription factor binding sites." Proc Int Conf Intell Syst Mol Biol **8**: 344-354.

Stangler, T., T. Tran, S. Hoffmann, H. Schmidt, E. Jonas and D. Willbold (2007). "Competitive displacement of full-length HIV-1 Nef from the Hck SH3 domain by a high-affinity artificial peptide." Biol Chem **388**(6): 611-615.

Stein, A. and P. Aloy (2008). "Contextual specificity in peptide-mediated protein interactions." PLoS One **3**(7): e2524.

Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen and C. von Mering (2015). "STRING v10: protein-protein interaction networks, integrated over the tree of life." Nucleic Acids Research **43**(D1): D447-D452.

Sztuba-Solinska, J., V. Stollar and J. J. Bujarski (2011). "Subgenomic messenger RNAs: Mastering regulation of (+)-strand RNA virus life cycle." Virology **412**(2): 245-255.

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

Teng, B., C. Zhao, X. Liu and Z. He (2015). "Network inference from AP-MS data: computational challenges and solutions." Brief Bioinform **16**(4): 658-674.

The UniProt, C. (2017). "UniProt: the universal protein knowledgebase." Nucleic Acids Res **45**(D1): D158-D169.

Thijs, G., K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouze and Y. Moreau (2002). "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes." J Comput Biol **9**(2): 447-464.

Tompa, P. (2012). "Intrinsically disordered proteins: a 10-year recap." Trends Biochem Sci **37**(12): 509-516.

Tompa, P. and P. Csermely (2004). "The role of structural disorder in the function of RNA and protein chaperones." FASEB J **18**(11): 1169-1175.

Tompa, P., N. E. Davey, T. J. Gibson and M. M. Babu (2014). "A million peptide motifs for the molecular biologist." Mol Cell **55**(2): 161-169.

Traweger, A., D. Fang, Y. C. Liu, W. Stelzhammer, I. A. Krizbai, F. Fresser, H. C. Bauer and H. Bauer (2002). "The tight junction-specific protein occludin is a functional target of the E3 ubiquitin-protein ligase itch." J Biol Chem **277**(12): 10201-10208.

Turner, B., S. Razick, A. L. Turinsky, J. Vlasblom, E. K. Crowdy, E. Cho, K. Morrison, I. M. Donaldson and S. J. Wodak (2010). "iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence." Database-the Journal of Biological Databases and Curation.

UniProt Consortium, T. (2018). "UniProt: the universal protein knowledgebase." Nucleic Acids Res **46**(5): 2699.

Uyar, B., R. J. Weatheritt, H. Dinkel, N. E. Davey and T. J. Gibson (2014). "Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer?" Mol Biosyst **10**(10): 2626-2642.

van der Lee, R., M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright and M. M. Babu (2014). "Classification of intrinsically disordered regions and proteins." Chem Rev **114**(13): 6589-6631.

van Helden, J., A. F. Rios and J. Collado-Vides (2000). "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads." Nucleic Acids Res **28**(8): 1808-1818.

Van Roey, K., B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson and N. E. Davey (2014). "Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation." Chemical Reviews **114**(13): 6733-6778.

Via, A., B. Uyar, C. Brun and A. Zanzoni (2015). "How pathogens use linear motifs to perturb host cell networks." Trends in Biochemical Sciences **40**(1): 36-48.

Wan, C., B. Borgeson, S. Phanse, F. Tu, K. Drew, G. Clark, X. Xiong, O. Kagan, J. Kwan, A. Bezginov, K. Chessman, S. Pal, G. Cromar, O. Papoulas, Z. Ni, D. R. Boutz, S. Stoilova, P. C. Havugimana, X. Guo, R. H. Malty, M. Sarov, J. Greenblatt, M. Babu, W. B. Derry, E. R. Tillier, J. B. Wallingford, J. Parkinson, E. M. Marcotte and A. Emili (2015). "Panorama of ancient metazoan macromolecular complexes." Nature **525**(7569): 339-344.

Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones (2004). "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life." J Mol Biol **337**(3): 635-645.

Weatheritt, R. J., N. E. Davey and T. J. Gibson (2012). "Linear motifs confer functional diversity onto splice variants." Nucleic Acids Res **40**(15): 7123-7131.

Weatheritt, R. J., P. Jehl, H. Dinkel and T. J. Gibson (2012). "iELM--a web server to explore short linear motif-mediated interactions." Nucleic Acids Res **40**(Web Server issue): W364-369.

Welch, E. J., B. W. Jones and J. D. Scott (2010). "Networking with AKAPs: context-dependent regulation of anchored enzymes." Mol Interv **10**(2): 86-97.

White, K. A., L. Enjuanes and B. Berkhout (2011). "RNA virus replication, transcription and recombination." RNA Biol **8**(2): 182-183.

Wickner, R. B. (1993). "Double-stranded RNA virus replication and packaging." J Biol Chem **268**(6): 3797-3800.

Wilhelm, M., J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J. H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber and B. Kuster (2014). "Mass-spectrometry-based draft of the human proteome." Nature **509**(7502): 582-587.

Wirblich, C., B. Bhattacharya and P. Roy (2006). "Nonstructural protein 3 of bluetongue virus assists virus release by recruiting ESCRT-I protein Tsg101." J Virol **80**(1): 460-473.

Xenarios, I., D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg (2000). "DIP: the database of interacting proteins." Nucleic Acids Res **28**(1): 289-291.

Xing, E. P., W. Wu, M. I. Jordan and R. M. Karp (2004). "Logos: a modular bayesian model for de novo motif detection." J Bioinform Comput Biol **2**(1): 127-154.

Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill and M. Vidal (2008). "High-quality binary protein interaction map of the yeast interactome network." Science **322**(5898): 104-110.

Zarrinpar, A., S. H. Park and W. A. Lim (2003). "Optimization of specificity in a cellular protein interaction network by negative selection." Nature **426**(6967): 676-680.

Zhang, A., T. M. Tessier, K. J. C. Galpin, C. R. King, S. F. Gameiro, W. W. Anderson, A. F. Yousef, W. T. Qin, S. S. C. Li and J. S. Mymryk (2018). "The Transcriptional Repressor BS69 is a Conserved Target of the E1A Proteins from Several Human Adenovirus Species." Viruses **10**(12).

Zhang, A. D., L. B. He and Y. P. Wang (2017). "Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions." Bmc Bioinformatics **18**.

Zhang, X. F., L. Ou-Yang, X. Hu and D. Q. Dai (2015). "Identifying binary protein-protein interactions from affinity purification mass spectrometry data." BMC Genomics **16**: 745.

Zhang, Y., H. Lin, Z. Yang and J. Wang (2015). "Integrating experimental and literature protein-protein interaction data for protein complex prediction." BMC Genomics **16 Suppl 2**: S4.