6-27-2022

In Silico Characterization of Protein-Protein Interactions Mediated by Short Linear Motifs

Heidy Elkhaligy
helkh002@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

IN SILICO CHARACTERIZATION OF PROTEIN-PROTEIN INTERACTIONS MEDIATED BY SHORT LINEAR MOTIFS

A thesis submitted in partial fulfillment of

the requirements for the degree of

MASTER OF SCIENCE

in

BIOLOGY

by

Heidy Elkhaligy

2022

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This thesis, written by Heidy Elkhaligy, and entitled In Silico Characterization of Protein-Protein Interactions Mediated by Short Linear Motifs, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

_____
Lidia Kos

_____
Yuk-Ching Tse-Dinh

_____
Prem Chapagain

_____
Jessica Siltberg-Liberles, Major Professor

Date of Defense: June 27, 2022

The thesis of Heidy Elkhaligy is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

I dedicate this thesis to my mom (1961-2013), family, friends, and everyone who supported me during my master's journey at Florida International University.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Jessica Siltberg-Liberles for her support, guidance, and mentoring during my master's studies. Without her help, I would not have been able to diversify my knowledge and learn about evolutionary biology and bioinformatics, improve my writing and presentation skills, attend international conferences to present my work, and learn how to communicate with other researchers.

I would also like to thank all my committee members, Dr. Lidia Kos, Dr. Yuk-Ching Tse-Dinh, and Dr. Prem Chapagain, for their continuous support and guidance and for their valuable conversations that have assisted me in my thesis research.

I would like to thank Dr. Yuan Liu for her guidance and support since I joined Florida International University and for constantly reminding me to look at the big picture to continue in my career. Thanks to Dr. Giri Narasimhan for his insightful discussions and allowing me to attend his RAPID project weekly meetings that have helped me learn more about molecular mimicry and different computational algorithms. Thanks to Dr. Tim Collins for his support as a graduate program director. Thanks to Alejandro J. Garcia Morales, the biology office specialist, for his patience and support in completing and submitting all the official documents in a timely manner.

Thanks to Christian A. Balbin for his support in helping me improve my coding skills, de-bugging the codes, and helping me in my research analysis. Thanks to graduate student Janelle Nunez-Castilla for her thoughtful discussions and for helping me learn new research techniques. Thanks to Jessica Gonzalez, Teresa Liberatore, and Daniel Morales, who helped me improve my teaching and writing. Thanks to all the lab members and the undergraduates who made my stay in the lab enjoyable.

I would like to thank my dad, Tarek Elkhaligy, and sister for being there when needed. I would like to thank my husband, Dr. Ahmed Seddek, and my kids, who have always been accommodating, understanding, encouraging, and supportive to me.

ABSTRACT OF THE THESIS

IN SILICO CHARACTERIZATION OF PROTEIN-PROTEIN INTERACTIONS MEDIATED BY SHORT LINEAR MOTIFS

by

Heidy Elkhaligy

Florida International University, 2022

Miami, Florida

Professor Jessica Siltberg-Liberles, Major Professor

Short linear motifs (SLiMs), often found in intrinsically disordered regions (IDPs), can initiate protein-protein interactions in eukaryotes. Although pathogens tend to have less disorder than eukaryotes, their proteins alter host cellular function through molecular mimicry of SLiMs. The first objective was to study sequence-based structure properties of viral SLiMs in the ELM database and the conservation of selected viral motifs involved in the virus life cycle. The second objective was to compare the structural features for SliMs in pathogens and eukaryotes in the ELM database. Our analysis showed that many viral SliMs are not found in IDPs, particularly glycosylation motifs. Moreover, analysis of disorder and secondary structure properties in the same motif from pathogens and eukaryotes shed light on similarities and differences in motif properties between pathogens and their eukaryotic equivalents. Our results indicate that the interaction mechanism may differ between pathogens and their eukaryotic hosts for the same motif.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF FIGURES

CHAPTER II

FIGURE PAGE

CHAPTER III

FIGURE PAGE

ABBREVIATIONS

ABBREVIATION	DEFINITION
PPI	Protein-protein interactions
ELM	Eukaryotic Linear Motifs
SLiMs	Short Linear Motifs
IDR	Intrinsically Disordered protein Region
IDP	Intrinsically Disordered Protein
PDB	Protein Data Bank
MIDS	Mean IUPRED2A Disorder Score
LIG	Ligand binding motifs
MOD	Post-translational modification motifs
TRG	Targeting motifs
DOC	Docking motifs
CLV	Cleavage sites motifs
DEG	Degradation motifs
ESCRT	Endosomal Sorting Complexes Required for Transport
PTM	Post-Translation Modifications
FMDV	Foot and Mouth Disease Virus
PI3K	Phosphatidyl-Inositol-3-Kinase
MAPK	Mitogen-Activated Protein Kinase
RBD	Receptor-Binding Domain
SARS-COV	Severe Acute Respiratory Syndrome Corona Virus
prM	precursor Membrane protein

HA	Hemagglutinin
HCMV	Human CytoMegalo Virus
STING	STimulator of INterferon Genes
MAVS	Mitochondrial AntiViral Signaling protein
TRIF	TIR domain-containing adaptor inducing IFN-β
IFN-3	Interferon regulatory factor 3
HPV	Human Papilloma Virus
Crb	Crumbs protein
PATJ	PALS1-Associated Tight Junction protein
MCCS	Mean Coil Confidence Score
mMIDS	mean MIDS
mMCCS	mean MCCS
WNV	West Nile Virus
Rb	Retinoblastoma protein
DSSP	Dictionary of Secondary Structure of Proteins

PREFACE

The following chapters have been published and are formatted according to journal specifications:

CHAPTER II

Elkhaligy, H., Balbin, C. A., Gonzalez, J. L., Liberatore, T., & Siltberg-Liberles, J. (2021). Dynamic, but Not Necessarily Disordered, Human-Virus Interactions Mediated through SLiMs in Viral Proteins. *Viruses*, *13*(12), 2369. https://doi.org/10.3390/v13122369

CHAPTER III

Elkhaligy, H., Balbin, C. A., & Siltberg-Liberles, J. (2022). Comparative Analysis of Structural Features in SLiMs from Eukaryotes, Bacteria, and Viruses with Importance for Host-Pathogen Interactions. *Pathogens*, *11*(5), 583. https://doi.org/10.3390/pathogens11050583.

CHAPTER I: INTRODUCTION

Protein structure determines function. However, proteins are allosteric, dynamic biomacromolecules that exist as conformational ensembles [1,2]. The energy landscape of proteins clearly illustrates the distinction between different protein conformations, which can range from being well folded (proteins that have a funnel-shaped energy landscape) to fully or partially unfolded proteins, known as intrinsically disordered proteins (IDPs), which have a more flattened energy landscape. A closer look at the bottom of the funnel-shaped energy landscape of folded proteins, especially those that exhibit allostery, reveals that most do not have one energy local minima. However, many have minimal energy differences between conformations, introducing the notion of folded proteins with IDRs. The minor energy difference in proteins with IDRs, allows them to endure conformational flexibility and undergo conformational changes [3]. Consequently, they can exist as ensembles of conformations that may have slightly or radically different functions. Conformationally flexible regions in proteins that may be unstructured under certain or all cellular conditions are called intrinsically disordered regions (IDR) [4–6]. IDRs can enable proteins to undergo large or small conformational changes [5,7,8], which may regulate protein function, promote interactions with other proteins, and more [5,9,10]. IDRs may also fold upon interacting with another biomolecule [5,11,12] or after being post-translationally modified [13].

IUPRED2A is a sequence-based predictor of protein disorder, which approximates the energy of each residue in a protein based on their type and the surrounding residues types within a window of 20 amino acids. The estimated energy is scaled to range from 0 to 1, where 0 indicates that the residue tends to be ordered (having a specific structure), and 1 indicates that the residue is highly likely to be disordered (lacking a definite

conformation). A cutoff value of 0.5 is usually used to differentiate between ordered and disordered residues [14]. The prediction performance of IUPRED2A was very similar to its earlier precedent IUPRED [14]. It was found that the accuracy of IUPRED using the default threshold of 0.5 against a dataset of disordered proteins from Disprot v.7 [15] was approximately 70%. However, another study using the same 0.5 threshold but a different dataset including disordered proteins from Disprot v.7 [15] and Protein Data Bank (PDB) [16] recorded a higher accuracy for both IUPRED and IUPRED2A of approximately 81% [17]. This discrepancies and low percent accuracy is due to the sparsity of the data that are used to train and test the models [18]. Although Disprot is the best available database for experimentally verified disordered proteins it is small, version 7 included about 800 proteins (~2100 segments) annotated to be disordered or comprise IDRs [15]. When considering protein structures found in PDB [16], unresolved residues are classified as disordered [19]. However, the presence of more than one PDB structure for the same protein region in varying context such as with an interacting protein or without can result in a mixture of missing and resolved structural regions and suggests a possible disorder to order transition [18,19]. Hence, the PDB structures in training and testing datasets may introduce bias and noise that can lead to misleading accuracies [18].

For IUPRED2A, IDRs in proteins can be classified as long or short disordered regions. Long disordered protein regions are extended unstructured regions within a protein, while short disordered regions are short disordered segments found within structured protein domains [14].

Protein-protein interactions (PPIs) are crucial for maintaining and regulating molecular functions in cellular organisms [20–22]. Alterations in the protein interaction network can lead to disease [20,22]. PPIs occurs when two or more proteins have a close physical interaction [23,24], such interaction can occur through structural interface complementarity. Structural interfaces are the primary determinant of the interaction between proteins [24,25]. The interaction between different protein surfaces can occur due to structural and sequence complementarity, or structural complementarity for the binding region, or through short amino acid sequences that are usually found in disordered protein regions which are known as short linear motifs (SLiMs) [24]. In this thesis, we will only focus on the SLiMs and study their sequence-based structural properties to explore the similarities and differences in SLiMs features between different taxonomic groups.

SLiMs are small amino acid stretches of 3 and 10 amino acids [26–29]. The presence of SLiMs in regions mostly lacking secondary structure or in disordered regions allows them to exist as flexible conformational ensembles that can participate in myriad protein interactions inside the cell [26,28,30,31]. While some SLiMs undergo conformational transitions with several structural representatives in the PDB, other SLiMs have no structural representative. Hence, a lower IUPRED cutoff value of 0.4 is used in previous studies and SLiM predictors to account for such motifs that have an increased tendency towards being ordered (can undergo disorder to order transition) [27,32]. However to represent all SLiMs similarly, prediction of secondary structure based on amino acid sequence is helpful using secondary structure predictors such as NetSurfP 2.0. NetSurfP 2.0 uses a deep neural network approach to predict the secondary structure of proteins by

comparing the input sequence to a reference clustered database of known PDB structures resulting in estimating the secondary structure of each residue with a calculated confidence value. NeSurfP 2.0 has an accuracy of ~85% using both mmseqs2 [33]and hhblits [34] methods of searching and clustering similar proteins to the query sequence against a Critical Assessment of protein Structure Prediction (CASP 12) datasets [35]. The regions that lack a specific secondary structure in the proteins are defined by NetSurfP 2.0 [36] as not being in an alpha helix or beta-sheets, and according to the simplified three-state secondary structure assignment of NetSurfP 2.0, they are classified as coil [36]. From hereafter, we will be using the three states structure assignment nomenclature assigned by NetSurfP 2.0.

Eukaryotic proteins include a higher percentage of disordered protein regions than bacteria and viruses [37,38]. Due to random mutational processes, bacterial and viral pathogens may acquire SLiMs that resemble one or more of their host protein SLiMs [39]. By mimicking host SLiMs, pathogens can hijack cellular functions and alter the signaling pathways of the host cell [40,41]. While most SLiMs interactions are transient [26,42,43], some pathogenic SLiMs were discovered to have a high affinity toward the interacting protein partner [13,44] and longer interaction time, which provides pathogens a better chance for altering the host cellular machinery [24]. Pathogens utilize motif mimicry to hinder the host immune response inside the cells[24,45], hijack replication machinery to allow their replication [46,47], facilitate their attachment [48,49], or egress from the host cells [50,51].

Data from experimentally verified SLiMs have been used to study SLiMs characteristics to create computational algorithms and tools [52–54] for SLiM identification. The prevalent database for manually-annotated experimentally verified SLiMs is the ELM database [53]. There are six categories of SLiMs in the ELM database: cleavage (CLV), degradation (DEG), docking (DOC), ligand binding (LIG), modification (MOD), and targeting sites (TRG). Most computational tools use a regular expression method for pattern search, including the ELM database webserver, to search for the motifs [52–54]. However, these patterns tolerate some variation at specific amino acid positions [28,29]. Given the short length of SLiMs and the variability at certain positions, the amino acid pattern may occur by chance. Thus, searching for patterns may find SLiMs that are not functional by chance numerous times in the protein sequence, which leads to a high false-positive rate [42,55].

Previous research showed that most eukaryotic SLiMs are found in disordered protein regions based on sequence-based disorder predictions such as IUPRED and PONDER [27]. Hence, excluding SLiMs in ordered protein regions was recommended to reduce false-positive SLiMs [55]. However, this method has been challenged as solely using disorder/order propensity will eliminate functional motifs found in ordered protein regions [56]. Another fundamental SLiM characteristic is the accessibility and secondary structure propensity. Via and coworkers investigated the structural characteristics of functional SLiMs and compared them to a random sample of proteins dataset [57]. In their analysis, accessible and coil regions were more enriched in the functional SLiMs than in the randomly sampled dataset. It was suggested that accessibility and being in the coil region could be used to find functional SLiMs. However, one drawback of their

method is that it will exclude buried SLiMs that are only accessible and functional due to allosteric structural changes [57].

Because eukaryotic proteins include more disordered regions than pathogens, such as bacteria and viruses, we hypothesize that SLiMs are less disordered in pathogens such as bacteria and viruses. SLiMs from different taxonomic groups may have different properties. Pathogenic proteins may be less allosteric with an increased shift towards a specific conformational ensemble that enables them to have a greater binding affinity to host proteins than the host SliMs they mimic. While some eukaryotic SliMs change their secondary structure conformation depending on the binding partner or the cellular context in a highly regulatory manner for the organism's fitness, pathogens hijack host cells and take over their cellular machinery have different functional constraints. This poses whether a SLiM that must form transient interactions to function properly in eukaryotes can form a more stable interaction in a mimicking pathogen protein with the host proteins.

In the second chapter, we performed sequence-based structure analysis on true positive SLiMs in the ELM database to study their disorder, accessibility, and secondary structure propensity. Selected examples from the viral motifs were further examined using multiple sequence alignment to explore the sequence conservation and sequence-based structure conservation of motifs in other homologous proteins. Such analysis would give us more insights into homologous proteins' structural features. In this chapter, we were able to show that viruses SLiMs have lower disorder content, which proves that previous studies about overall disorder content in viral proteins are low. This indicates

that there might be some differences in SLiMs characteristics between different taxonomic groups.

In the third chapter, a comparative analysis of the sequence-based structural features of the true positive SLiMs in the ELM database was performed to explore the differences or similarities between taxonomic groups and different ELM types. This computational analysis is the first extensive analysis between different taxonomic groups. Our analysis revealed sequence-based structural differences between taxonomic groups and between different specific ELM types. Moreover, the analysis revealed that some ELM types shared between pathogens and eukaryotes have different properties that may enhance the pathogenicity or virulence of the pathogens.

SIGNIFICANCE

To date, we do not have a complete understanding of all the SLiMs' attributes or how they vary between taxonomic groups such as eukaryotes, bacteria, and viruses. With the increased amount of annotated SLiMs data, more knowledge can be derived from these experimentally verified SLiMs to help discover and identify functional motifs and differences across taxonomic groups. The improved discovery of functional SLiMs and how SLiMs attributes may differ for hosts, and their pathogens will provide the scientific community with a better understanding of protein interactions and host-pathogen protein interactions.

REFERENCES

1. Gunasekaran, K.; Ma, B.; Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins Struct. Funct. Bioinforma.* **2004**, *57*, 433–443, doi:10.1002/PROT.20232.

2. Boehr, D.D.; Nussinov, R.; Wright, P.E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5*, 789–796, doi:10.1038/nchembio.232.

3. Burger, V.M.; Gurry, T.; Stultz, C.M. Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polym.* **2014**, *6*, 2684–2719, doi:10.3390/POLYM6102684.

4. Mittag, T.; Kay, L.E.; Forman-Kaya, J.D. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* **2010**, *23*, 105–116, doi:10.1002/JMR.961.

5. Van Der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; et al. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114*, 6589–6631, doi:10.1021/CR400525M.

6. Gao, C.; Ma, C.; Wang, H.; Zhong, H.; Zang, J.; Zhong, R.; He, F.; Yang, D. Intrinsic disorder in protein domains contributes to both organism complexity and clade-specific functions. *Sci. Reports* **2021**, *11*, 1–18, doi:10.1038/s41598-021-82656-9.

7. Radivojac, P.; Obradovic, Z.; Smith, D.K.; Zhu, G.; Vucetic, S.; Brown, C.J.; Lawson, J.D.; Dunker, A.K. Protein flexibility and intrinsic disorder. *Protein Sci.* **2004**, *13*, 71, doi:10.1110/PS.03128904.

8. Uversky, V.N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Front. Phys.* **2019**, *0*, 10, doi:10.3389/FPHY.2019.00010.

9. Yang, L.Q.; Sang, P.; Tao, Y.; Fu, Y.X.; Zhang, K.Q.; Xie, Y.H.; Liu, S.Q. Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *J. Biomol. Struct. Dyn.* **2014**, *32*, 372, doi:10.1080/07391102.2013.770372.

10. Babu, M.M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **2016**, *44*, 1185, doi:10.1042/BST20160172.

11. Babu, M.M.; Kriwacki, R.W.; Pappu, R. V. Versatility from protein disorder. *Science.* **2012**, *337*, 1460–1461.

12. Radivojac, P.; Iakoucheva, L.M.; Oldfield, C.J.; Obradovic, Z.; Uversky, V.N.; Dunker, A.K. Intrinsic Disorder and Functional Proteomics. *Biophys. J.* **2007**, *92*, 1439–1456, doi:10.1529/BIOPHYSJ.106.094045.

13. Van Roey, K.; Uyar, B.; Weatheritt, R.J.; Dinkel, H.; Seiler, M.; Budd, A.; Gibson, T.J.; Davey, N.E. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chem. Rev.* **2014**, *114*, 6733–6778, doi:10.1021/CR400585Q.

14. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337, doi:10.1093/nar/gky384.

15. Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C.J.; Aspromonte, M.C.; Davey, N.E.; Davidović, R.; Dosztányi, Z.; et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **2017**, *45*, D219–D227, doi:10.1093/NAR/GKW1056.

16. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C.H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J.M.; et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2021**, *49*, D437–D451, doi:10.1093/NAR/GKAA1038.

17. Zhao, B.; Xue, B. Decision-Tree Based Meta-Strategy Improved Accuracy of Disorder Prediction and Identified Novel Disordered Residues Inside Binding Motifs. *Int. J. Mol. Sci.* **2018**, *19*, doi:10.3390/IJMS19103052.

18. DeForte, S.; Uversky, V.N. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. *Protein Sci.* **2016**, *25*, 676, doi:10.1002/PRO.2864.

19. Gall, T. Le; Romero, P.R.; Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Intrinsic disorder in the protein data bank. *J. Biomol. Struct. Dyn.* **2007**, *24*, 325–341, doi:10.1080/07391102.2007.10507123.

20. Braun, P.; Gingras, A.C. History of protein–protein interactions: From egg-white to complex networks. *Proteomics* **2012**, *12*, 1478–1498, doi:10.1002/PMIC.201100563.

21. Peng, X.; Wang, J.; Peng, W.; Wu, F.X.; Pan, Y. Protein–protein interactions: detection, reliability assessment and applications. *Brief. Bioinform.* **2017**, *18*, 798–819, doi:10.1093/BIB/BBW066.

22. Rao, V.S.; Srinivas, K.; Sujini, G.N.; Kumar, G.N.S. Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics* **2014**, *2014*, 1–12, doi:10.1155/2014/147648.

23. De Las Rivas, J.; Fontanillo, C. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* **2010**, *6*, 1–8, doi:10.1371/journal.pcbi.1000807.

24. Guven-Maiorov, E.; Tsai, C.J.; Nussinov, R. Pathogen mimicry of host protein-protein interfaces modulates immunity. *Semin. Cell Dev. Biol.* **2016**, *58*, 136–145, doi:10.1016/J.SEMCDB.2016.06.004.

25. Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R. Principles of Protein−Protein Interactions: What are the Preferred Ways For Proteins To Interact? *Chem. Rev.* **2008**, *108*, 1225–1244, doi:10.1021/CR040409X.

26. Davey, N.E.; Van Roey, K.; Weatheritt, R.J.; Toedt, G.; Uyar, B.; Altenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T.J. Attributes of short linear motifs. *Mol. Biosyst.* **2012**, *8*, 268–281, doi:10.1039/c1mb05231d.

27. Fuxreiter, M.; Tompa, P.; Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **2007**, *23*, 950–956, doi:10.1093/bioinformatics/btm035.

28. Hagai, T.; Azia, A.; Babu, M.M.; Andino, R. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell Rep.* **2014**, *7*, 1729–1739, doi:10.1016/j.celrep.2014.04.052.

29. Sobhy, H. A review of functional motifs utilized by viruses. *Proteomes* **2016**, *4*, doi:10.3390/proteomes4010003.

30. Cumberworth, A.; Lamour, G.; Babu, M.M.; Gsponer, J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochem. J.* **2013**, *454*, 361–369, doi:10.1042/BJ20130545.

31. O'Shea, C.; Staby, L.; Bendsen, S.K.; Tidemand, F.G.; Redsted, A.; Willemoës, M.; Kragelund, B.B.; Skriver, K. Structures and Short Linear Motif of Disordered Transcription Factor Regions Provide Clues to the Interactome of the Cellular Hub Protein Radical-induced Cell Death1. *J. Biol. Chem.* **2017**, *292*, 512, doi:10.1074/JBC.M116.753426.

32. Dosztányi, Z. Prediction of protein disorder based on IUPred. *Protein Sci.* **2018**, *27*, 331–340, doi:10.1002/PRO.3334.

33. Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028.

34. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2011**, *9*, 173–175, doi:10.1038/nmeth.1818.

35. Schaarschmidt, J.; Monastyrskyy, B.; Kryshtafovych, A.; Bonvin, A.M.J.J. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins Struct. Funct. Bioinforma.* **2018**, *86*, 51–66, doi:10.1002/PROT.25407.

36. Klausen, M.S.; Jespersen, M.C.; Nielsen, H.; Jensen, K.K.; Jurtz, V.I.; Sønderby, C.K.; Sommer, M.O.A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinforma.* **2019**, *87*, 520–527, doi:10.1002/prot.25674.

37. Kastano, K.; Erdős, G.; Mier, P.; Alanis-Lobato, G.; Promponas, V.J.; Dosztányi, Z.; Andrade-Navarro, M.A. Evolutionary Study of Disorder in Protein Sequences. *Biomolecules* **2020**, *10*, 1–17, doi:10.3390/BIOM10101413.

38. Peng, Z.; Yan, J.; Fan, X.; Mizianty, M.J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V.N.; Kurgan, L. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **2015**, *72*, 137–151, doi:10.1007/S00018-014-1661-9.

39. Franzosa, E.A.; Xia, Y. Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci.* **2011**, *108*, 10538–10543, doi:10.1073/PNAS.1101440108.

40. Davey, N.E.; Travé, G.; Gibson, T.J. How viruses hijack cell regulation. *Trends Biochem. Sci.* **2011**, *36*, 159–169, doi:10.1016/J.TIBS.2010.10.002.

41. Sámano-Sánchez, H.; Gibson, T.J. Mimicry of Short Linear Motifs by Bacterial Pathogens: A Drugging Opportunity. *Trends Biochem. Sci.* **2020**, *45*, 526–544, doi:10.1016/J.TIBS.2020.03.003.

42. Hraber, P.; O'Maille, P.E.; Silberfarb, A.; Davis-Anderson, K.; Generous, N.; McMahon, B.H.; Fair, J.M. Resources to Discover and Use Short Linear Motifs in Viral Proteins. *Trends Biotechnol.* **2020**, *38*, 113–127.

43. Davey, N.E.; Cyert, M.S.; Moses, A.M. Short linear motifs – ex nihilo evolution of protein regulation. *Cell Commun. Signal.* **2015**, *13*, 1–15, doi:10.1186/S12964-015-0120-Z.

44. Palopoli, N.; Foutel, N.S.G.; Gibson, T.J.; Chemes, L.B. Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity. *Protein Eng. Des. Sel.* **2018**, *31*, 69–77, doi:10.1093/PROTEIN/GZX068.

45. Alcami, A.; Koszinowski, U.H.; Alcami, A.; Koszinowski, U.H. Viral mechanisms of immune evasion. *Trends Microbiol.* **2000**, *8*, 410–418, doi:10.1016/S0966-842X(00)01830-8.

46. Finnen, R.L.; Pangka, K.R.; Banfield, B.W. Herpes Simplex Virus 2 Infection Impacts Stress Granule Accumulation. *J. Virol.* **2012**, *86*, 8119, doi:10.1128/JVI.00313-12.

47. Felsani, A.; Mileo, A.M.; Paggi, M.G. Retinoblastoma family proteins as key targets of the small DNA virus oncoproteins. *Oncogene* **2006**, *25*, 5277–5285.

48. Hussein, H.A.M.; Walker, L.R.; Abdel-Raouf, U.M.; Desouky, S.A.; Montasser, A.K.M.; Akula, S.M. Beyond RGD: virus interactions with integrins. *Arch. Virol.* **2015**, *160*, 2669, doi:10.1007/S00705-015-2579-8.

49. Barden, S.; Lange, S.; Tegtmeyer, N.; Conradi, J.; Sewald, N.; Backert, S.; Niemann, H.H. A helical RGD motif promoting cell adhesion: crystal structures of the Helicobacter pylori type IV secretion system pilus protein CagL. *Structure* **2013**, *21*, 1931–1941, doi:10.1016/J.STR.2013.08.018.

50. Welker, L.; Paillart, J.-C.; Bernacchi, S. Importance of Viral Late Domains in Budding and Release of Enveloped RNA Viruses. *Viruses* **2021**, *13*, doi:10.3390/V13081559.

51. Rose, K.M. When in need of an ESCRT: The nature of virus assembly sites suggests mechanistic parallels between nuclear virus egress and retroviral budding. *Viruses* **2021**, *13*, doi:10.3390/v13061138.

52. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, doi:10.1093/nar/gkp335.

53. Kumar, M.; Gouw, M.; Michael, S.; Sámano-Sánchez, H.; Pancsa, R.; Glavina, J.; Diakogianni, A.; Valverde, J.A.; Bukirova, D.; Čalyševa, J.; et al. ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* **2020**, *48*, D296–D306, doi:10.1093/NAR/GKZ1030.

54. Krystkowiak, I.; Davey, N.E. SLiMSearch: A framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Res.* **2017**, *45*, W464–W469, doi:10.1093/nar/gkx238.

55. Gould, C.M.; Diella, F.; Via, A.; Puntervoll, P.; Gemünd, C.; Chabanis-Davidson, S.; Michael, S.; Sayadi, A.; Bryne, J.C.; Chica, C.; et al. ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.* **2010**, *38*, D167–D180, doi:10.1093/NAR/GKP1016.

56. Gibson, T.J.; Dinkel, H.; Van Roey, K.; Diella, F. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun. Signal.* **2015**, *13*, doi:10.1186/S12964-015-0121-Y.

57. Via, A.; Gould, C.M.; Gemünd, C.; Gibson, T.J.; Helmer-Citterich, M. A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics* **2009**, *10*, 1–17, doi:10.1186/1471-2105-10-351.

CHAPTER II: DYNAMIC, BUT NOT NECESSARILY DISORDERED, HUMAN-VIRUS INTERACTIONS MEDIATED THROUGH SLIMS IN VIRAL PROTEINS

ABSTRACT

Most viruses have small genomes that encode proteins needed to perform essential enzymatic functions. Across virus families, primary enzyme functions are under functional constraint; however, secondary functions mediated by exposed protein surfaces that promote interactions with the host proteins may be less constrained. Viruses often form transient interactions with host proteins through conformationally flexible interfaces. Exposed flexible amino acid residues are known to evolve rapidly, suggesting that secondary functions may generate diverse interaction potentials between viruses within the same viral family. One mechanism of interaction is viral mimicry through short linear motifs (SLiMs) that act as functional signatures in host proteins. Viral SLiMs display specific patterns of adjacent amino acids that resemble their host SLiMs and may occur by chance numerous times in viral proteins due to mutational and selective processes. Through mimicry of SLiMs in the host cell proteome, viruses can interfere with the protein interaction network of the host and utilize the host-cell machinery to their benefit. The overlap between rapidly evolving protein regions and the location of functionally critical SLiMs suggest that these motifs and their functional potential may be rapidly rewired causing variation in pathogenicity, infectivity, and virulence of related viruses. The following review provides an overview of known viral SLiMs with select examples of their role in the life cycle of a virus, and a discussion of the structural properties of experimentally validated SLiMs highlighting that a large portion of known viral SLiMs are devoid of predicted intrinsic disorder based on the viral SLiMs from the ELM database.

1. Introduction

Viruses are pathogens that cannot thrive outside a host [1,2]. Depending on the viral family, genomic information is encoded in either positive or negative single-stranded or double-stranded DNA or RNA. The genomic material is typically small, ranging from a few kb to over 1000 kb [3]. Viruses exploit host cell proteins to complete their life cycle: attachment, penetration, uncoating, replication and protein expression, assembly, and egress from the infected cell [1]. The viral genome is translated into structural proteins, nonstructural proteins, and sometimes accessory proteins. Structural proteins encapsulate the newly formed virus genome inside the host cell and provide the virion its shape. Non-structural proteins (nsps) typically make up the genome replication complex and include a polymerase that is dedicated to replicating the viral genome. Further, nsps partake in protein processing and may also perform secondary functions involved in impacting immune regulation and antiviral response. Accessory proteins are mainly regulatory proteins primarily involved in modulating host cell gene expression, inducing apoptosis, or affecting the viral rate of replication [4].

Viruses have high mutation rates [5], which is particularly true with regard to RNA viruses [6]. The fitness of RNA viruses depends on their RNA polymerases to replicate the viral genome with low fidelity [7,8]. While the primary enzymatic functions typically are under selective constraint, rapidly evolving amino acid residues are often located in conformationally flexible regions on the surface of the protein. Surfaces of viral proteins are major contact points to their hosts. Through interface mimicry, where a part of a viral protein surface resembles a host protein, the virus can interfere with protein-protein

networks of the host protein [9]. The presence of short linear motifs (SLiMs) that act as functional signatures in proteins are important for understanding protein-protein interactions in an organism. Identification of a SLiM from a host species in a viral protein suggests interface mimicry that may disrupt endogenous protein-protein interactions. Many host-virus mimicry-driven interactions are transient [10] and depend on the proteomic context of the host cell. Consequently, exogeneous interactions may give rise to complex diversity in viral virulence, pathogenicity, and transmissibility not only between different host species, but also within the same host species.

1.1. Short Linear Motifs

Eukaryotic Linear Motifs (ELMs) (a.k.a. SLiMs) are small segments of proteins, usually 3 to 10 amino acids long with a specific cellular function [11,12]. Given the linear sequence pattern that composes a SLiM, some positions in a SLiM can withstand various amino acid substitutions without affecting functionality, while an amino acid substitution at a different, critical position can eliminate all functionality. To represent sequence variation, SLiMs are described by regular expressions using the one-letter amino acid abbreviations [13]. Virus proteins that display SLiMs can perform molecular interactions with host proteins in a similar manner as the host protein it mimics [11]. SLiMs that occur in humans may also occur by chance in viral proteins due to convergent evolution [10]. SLiMs can occur in highly conserved protein regions or regions with a high evolutionary rate of amino acid substitution. The presence of conserved motifs within the same virus family suggests the existence of functionally important virus-host protein interactions. Conversely, the presence of rapidly evolving motifs can enable the emergence of new protein-protein interactions within different hosts [11,14].

1.2. SLiMs in Intrinsically Disordered Protein Regions

Intrinsically disordered regions (IDRs) lack a specific folded structure (order) and harbor high conformational plasticity [15]. Linear motifs from eukaryotes were found to be predominantly disordered based on prediction of intrinsic disorder [16]. Viral motifs within intrinsically disordered protein regions (IDRs) can enable viral-host protein interactions [2,11,12]. IDRs provide SLiMs malleability to interact with various target proteins and to acquire different transient secondary structures that facilitate SLiM interaction with another protein [11,15,17–19]. The plasticity of SLiMs has been proposed to impact viral phenotypic traits such as tropism and virulence [20].

A positive correlation between disorder content and the occurrence of linear motifs has been shown [11]. However, disorder content has been found to vary greatly between virus families and coronaviruses have among the least [21]. Proteome-wide evolutionary studies of coronaviruses revealed a highly disordered nucleocapsid protein while the other proteins had almost no disorder [22]. Yet, from the large SARS-CoV-2 data that has been accumulating over the last two years, it is apparent that coronaviruses such as SARS-CoV-2 perform a wealth of interactions with proteins in its human host despite a low predicted intrinsic disorder content.

2. Methods Used in the Discovery of SLiMs

2.1. Experimental Procedures

SLiMs are typically involved in transient protein-protein interactions (PPIs) with a low affinity towards the interacting protein [23,24]. Thus, mass spectroscopic analysis of PPIs might be unable to detect the SLiMs' temporary interactions in their normal mode; more specific optimizations are needed [25]. Other methods that have been proposed for

the discovery and investigation of SLiM interactions are peptide phage display and large-scale proteomic peptide phage display [26]. Phage display may be coupled with site-directed mutagenesis to verify the interacting pattern. One major disadvantage of the experimental methods exploring SLiMs on the peptide level is that the actual interaction inside the cell might not be properly portrayed due to the absence of post-translational protein modifications that are critical for the functionality of the SLiM [26].

2.2. Computational Approaches

Data from experimentally verified SLiMs can be used to make predictors or search functions for similar motifs. Various webservers with databases of linear motifs provide a search function for similar motifs using regular expression patterns (regex). According to the ELM database [27], the regex pattern symbols used are as follows: dot "." means that this position permits the presence of any amino acid which can be symbolized by "x" as well, square brackets "[ ]" mean any listed amino acid is accepted at that position, caret sign inside a square bracket "[^]" means that any following amino acid is not allowed in this site, curly brackets "{ }" specify the count or range of accepted amino acids at specific position in the pattern, dollar sign "$" indicates the C-terminal end of the protein sequence, caret sign "^" indicates the N-terminal end of the protein, question mark "?" indicates one optional amino acid (one or none), asterisk "*" specifies any number of optional amino acids is allowed (zero or more), plus sign "+" indicates one or more amino acids are accepted, pipe "|" separates and suggests an alternative amino acid pattern for the motif, and parentheses "( )" can either be used to group pieces of pattern or to indicate an important amino acid site such as covalently modified amino acids.

The ELM database is the prevalent resource for SLiMs. This database provides experimentally verified SLiMs classified as true positives [27]. SLiMs are categorized by function as either cleavage, degradation, docking, ligand binding, modification, or targeting sites [27]. Cleavage sites (CLV) are patterns identified by different proteolytic enzymes. Degradation sites (DEG) are sequences recognized for ubiquitination to allow subsequent protein breakdown. Docking sites (DOC) are involved in regulating protein interaction. Ligand binding sites (LIG) participate in protein-protein interactions. Modification sites (MOD) include amino acid patterns predicted to undergo post-translational modification. Targeting sites (TRG) act as signals for translocation of proteins [12,27].

Other resources are available such as SLiMSearch and MEME suite. SLiMSearch is a webserver that allows the user to input a regex pattern or motif consensus sequence and then choose the species where the motif is predicted to be found, along with other filtration options such as disorder cutoff value. The results provide proteins that potentially include the input motif with their predicted conservation score, relative disorder score, accessibility prediction, PTM predictions at the motif site, the presence of known, mutational SNPs in that region, and more data that can allow the user to filter the results based on their needs [28]. MEME suite includes many tools and pipelines for *de novo* motif discovery and searching for known motif patterns in your input dataset as well as performing enrichment analyses and more [29].

A critical challenge for the computational techniques is their high false-positive rate [12,30,31]. Filtration to reduce false positives such as making sure the SLiM is in a disordered region is commonly recommended and integrated in some tools such as SLiMSuite [32] and IUPRED3 [33].

3. Are Viral SLiMs Disordered?

SLiMs from the ELM database were shown to be disordered using mean IUPRED2A disorder scores (MIDS) [16,34]. IUPRED2A predicts a disorder score for amino acid residues in proteins [35,36]. If the score for a residue is greater than 0.5, that residue is predicted to be disordered. However, a cutoff of 0.4 has been shown to be in greater agreement with experimentally confirmed intrinsic disorder [16]. Considering a 0.4 cutoff, 78% [16] and 71% [34] of all motifs were found to have a MIDS above 0.4 indicating that some residues in some motifs are likely ordered.

To the best of our knowledge no study has investigated the viral SLiMs separately. With the large variation in disordered content in virus families [21], we were curious about the disorder content in viral SLiMs. To investigate the disorder content of linear motifs from viruses that are known to interact with host proteins, we undertook a brief study in that respect. We downloaded the FASTA sequences for all 260 viral SLiMs classified as true positives from the ELM database [27]. This dataset contains 131 LIG, 65 MOD, 38 TRG, 11 DOC, 11 CLV, and 4 DEG viral SLiM sites. For each sequence, we extracted the motif plus 50 flanking amino acid residues on the N-terminal and C-terminal sides, respectively. For sequences where the motif was located closer than 50 amino acid residues from a terminal, all residues towards that terminal were included. The resulting sequence fragments were used to predict intrinsic disorder with IUPRED2A

[35,36] (default settings). The predicted state was mapped to the corresponding position in each sequence using an IUPRED2A disorder score cutoff of 0.4 (and 0.5 separately) to infer disorder or order. Thereafter, the percentage of disordered residues for each motif region was calculated. We also calculated MIDS per motif.

We found that 38% of the viral motifs are completely disordered and another 38% are completely ordered based on IUPRED2A disorder scores with cutoff = 0.5. For the remaining motifs, disorder content varies (Figure 1a). Based on IUPRED2A disorder scores with cutoff = 0.4, 66 motifs (25%) are 100% ordered and 143 motifs (55%) are 100% disordered (Figure 1b). The predominant motif classes vary between the fully ordered and the fully disordered motifs. Of the fully disordered motifs, the predominant motifs are LIG (63%) and TRG (18%). Of the fully ordered motifs, the predominant motif classes are MOD (62%) and LIG (15%). MIDS revealed that >36% of all viral motifs had an average score below 0.4 (Figure 1c). These results suggest that screening for only disordered motifs may exclude a large portion of functional viral motifs and especially sites that undergo post-translational modification.

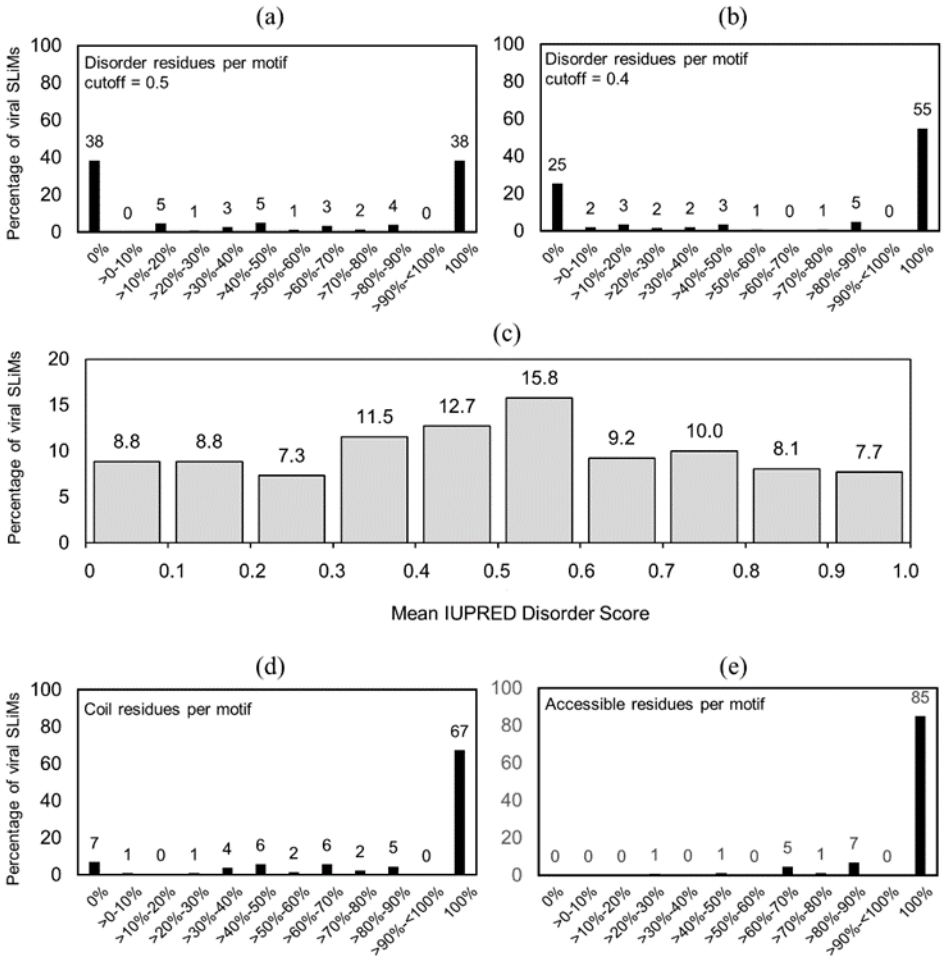

Figure 1. Predicted structural features of 260 viral SLiMs from the ELM database. The percentage of viral motifs with a certain disorder content as inferred from IUPRED2A prediction using a cutoff of (a) 0.5 and (b) 0.4. (c) The percentage of viral motifs with a certain Mean IUPRED2A Disorder Score (MIDS). The percentage of viral motifs with a certain (d) secondary structure (coil) and (e) surface accessibility content as inferred from NetSurfP-2.0 prediction. The percentages shown are approximate; rounded to the nearest whole number for a, b, d, and e, and to the nearest tenth for c. See also Table S1.

Further, we also predicted surface accessibility and secondary structure for the 260 viral motifs with NetSurfP-2.0 [37] with default settings. The NetSurfP-2.0 predictions were used to infer “not alpha helix or beta strand” as coil and surface accessibility for each residue in the motif. Thereafter, the fraction of coil and surface accessible residues

for each motif region was calculated. Most motifs are as expected surface accessible and tend to lack secondary structure. From the 260 viral motifs, 175 motifs (67%) are 100% coil, and 221 motifs (85%) are completely surface accessible (Figure 1d,e).

Based on prediction of disorder, surface accessibility, and secondary structure, our results suggest that a large portion of the true positive viral SLiMs are not disordered but a clear majority are in a coil conformation and an even stronger signal is seen from prediction of surface accessibility. Ultimately, these results, based on predictions of a limited set of viral linear motifs known to interact with host proteins, imply that viral SLiMs may not be as disordered as their analogous counterparts in eukaryotes. Further analyses are warranted to establish how disorder content varies for the same SLiM in a virus and its host. Here, we show selected examples of SLiMs that illustrate how disorder, surface accessibility, and secondary structure may vary across related viruses.

4. Select Viral SLiMs Involved in the Viral Life Cycle

The viral life cycle can be divided into events that occur outside the cell and inside the infected cell. In a general viral lytic cycle (Figure 2), the virus must first attach and fuse to the outside of the host cell before it can enter the cell. Then, the virus gets encapsulated or penetrates the cell membrane. Next, the virus starts the process of replication and translating its proteins to produce more viruses that are capable of infecting other neighboring cells. At this step, viral proteins hover inside the cell and migrate to several subcellular locations. As for host proteins, the presence of SLiMs in viruses may aid in the shuttling of viral proteins to different cellular compartments, where they can interact with various host proteins [27]. Finally, the virus particles are assembled, followed by viral exit from the infected cell [1].

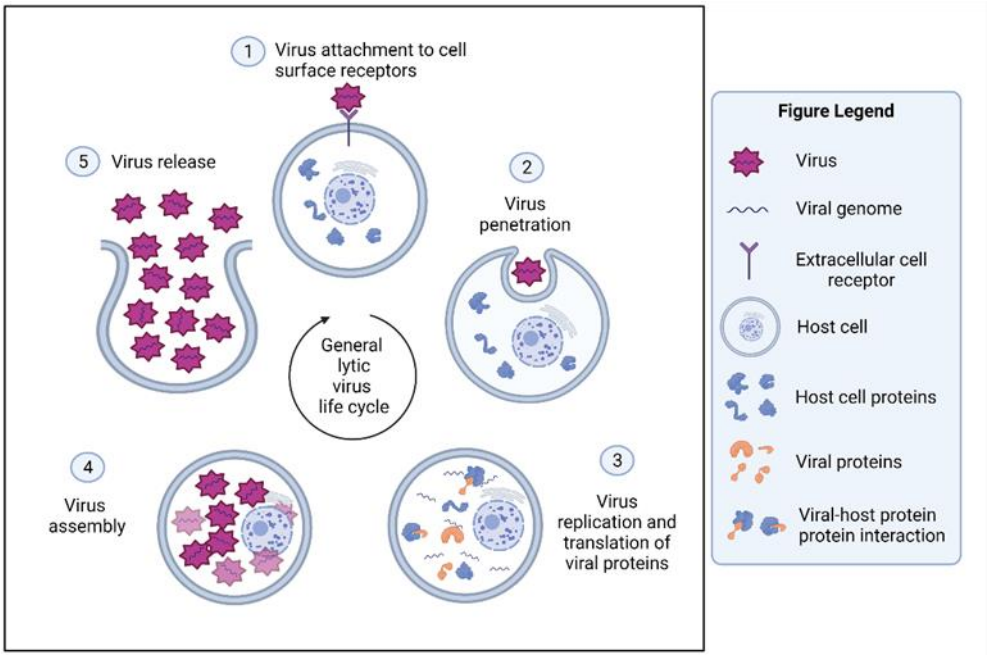

Figure 2. The general lytic virus life cycle inside the cells. (1) The virion attaches to the cell surface receptors. (2) The penetration of the virus through endocytosis to the infected cell. (3) The replicated genome and translated viral proteins inside the cell. (4) The newly assembled viruses inside the cell. (5) The cell lysis and release of new viruses from the infected cell. Created with BioRender.com (accessed on 30 October 2021)

4.1. SLiMs and Viral Cell Invasion through Cellular Attachment, Entry, and Fusion

4.1.1 RGD Motif, Integrin-Binding, and Attachment

The existence of specific motifs can enhance the ability of a virus to attach to the host cell receptors. For instance, the presence of the RGD pattern in virus envelope or membrane proteins, such as for Foot and Mouth disease virus (FMDV) [38] and Epstein-Barr virus [39], may promote viral fusion with host cells by facilitating the interaction with the integrin cell surface receptors [40]. Integrin receptors are transmembrane receptors that are involved in various signaling pathways including cellular communication with the surrounding environment. Several cell types, such as pneumocytes, endothelial cells, and platelets, express integrin transmembrane receptors.

When transmembrane integrin receptors recognize and bind to a pattern of RGD amino acids present on extracellular proteins, it can result in activation or inhibition of the integrin receptor's signaling pathways [41]. RGD integrin-binding activates clathrin-mediated endocytosis in adenoviruses and promotes virus entry into cells, triggering the phosphatidylinositol-3-kinase (PI3K) and mitogen-activated protein kinase (MAPK) pathways inside the infected cells. PI3K and MAPK are critical signaling pathways that control cell survival and proliferation [42].

The spike receptor-binding domain (RBD) from SARS-CoV-2 has an RGD motif that thus far is not found in other closely related coronaviruses [43]. The motif shows a degree of structural resemblance to other experimentally confirmed RGD-containing ligands and proteins that can bind to integrin receptors. Although the motif is not completely solvent accessible, it is located near a disordered protein region which may expose the RGD motif in a subset of the conformational ensemble enough to enable integrin binding under some conditions [44]. It has been speculated that the RGD motif could (1) promote the entry of SARS-CoV-2 into cells not expressing the primary SARS-CoV-2 receptor, the ACE2 receptor [45], and (2) affect the infectivity of the SARS-CoV-2 virus [43,44] due to the conformational flexibility surrounding the motif [44].

4.1.2 Furin Cleavage Motif Role in Viral Entry

To enhance cell entry, numerous viruses use a motif of the furin recognition pattern. Furin is a ubiquitously expressed protease [46] that promotes splitting and activation of various human extracellular proteins including hormones, growth factors, cellular receptors, adhesion molecules, and more [47]. Furin recognition patterns, R.[RK]R., where furin cleaves the protein after the last Arginine (R) in the pattern, have

been confirmed experimentally in HIV-1 [48], Coronaviruses [49], Flaviviruses [50], and other viruses (discussed in [47]), and in some bacterial toxins such as Anthrax toxin [51] and Diphtheria toxin [52].

In viruses, furin cleavage can lead to activation and facilitation of the viral fusion to cellular receptors and cell entry [53,54]. In Flaviviruses, furin proteolysis of precursor membrane (prM) protein is required to develop mature viruses [55]. In Orthomyxoviruses, such as influenza viruses, hemagglutinin (HA) glycoprotein cleavage leads to activation of the virus by unveiling the fusion peptide responsible for cell fusion and entry [56]. HA cleavage in avian influenza viruses was found responsible for the increased pathogenicity [53].

The conservation of sequence, disorder, and accessibility of the furin cleavage motif in HIV-1 [48] is high across sequences of HIV-1 envelope homologs suggesting a conserved function (Figure 3).

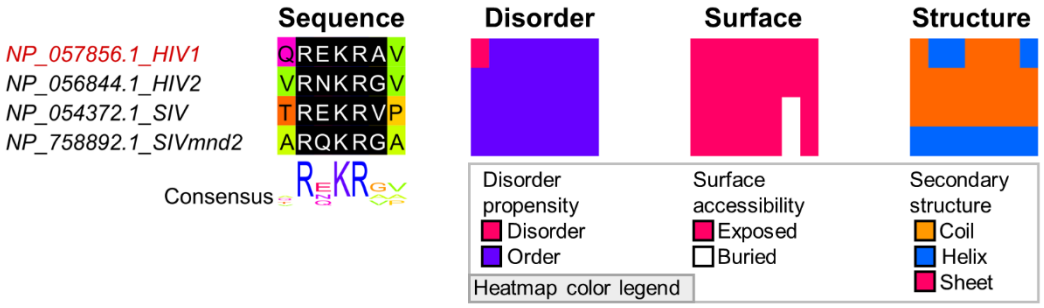

Figure 3. The furin cleavage site in the envelope glycoprotein from HIV. Sequences were identified with BLAST using the envelope protein (accession: NP_057856.1) from HIV-1 as query. Sequence names shown in red represents true positive instances from the ELM database [27]. The multiple sequence alignment (MSA) was built with MAFFT+L-INS-i [57] in Jalview [58]. The regular expression pattern R.[RK]R. from motif CLV_PCSK_FUR_1 in the ELM database [27] was identified using Find in Jalview, shown in black with white text. The region shown under Sequence shows the amino acids that corresponds to the true positive motif from ENV_HIV1 plus one additional site on each side. The three additional heatmaps display the same region of the alignment

colored by property. The heatmap for Disorder propensity displays disordered (magenta) or ordered (purple) residues based on IUPRED2A prediction with cutoff = 0.4 [35,36,59]. Heatmaps for (1) Surface accessibility displays surface exposed (magenta) and buried (white) residues and (2) Secondary structure displays coil (orange) and secondary structure (helix: blue, strand: magenta) based on NetSurfP-2.0 predictions.

In SARS-CoV-2, an additional furin cleavage site, absent in other closely related coronaviruses, was detected in the spike protein using sequence-based methods and it was suggested to be one of the principal causes of its pathogenicity [60]. Later, it was shown that while furin plays a role in successful SARS-CoV-2 infection, it is not critical for infection [61]. Further, other coronaviruses such as SARS-CoV also include furin recognition sites in nearby regions, and some of them were experimentally verified to be functional [62], which suggests that the exact position is not always critical for an analogous function.

4.2. SLiMs Influencing Viral Cell Replication

4.2.1 Retinoblastoma-Binding LxCxE Motif

After viruses invade the host cell, the viral genome is unpacked, and genome replication is initiated. For viral replication to occur, tampering with the host cell machinery is often achieved by promoting degradation of host proteolytic enzymes responsible for breaking down virus proteins, inhibiting degradation of host proteins essential for virus survival, and altering the host cell cycle by forcing the cell to the S phase [2]. Viruses may induce host cells to the S phase to facilitate their replication through the RB-binding LxCxE motif. Retinoblastoma proteins (RBs) are tumor suppressor proteins that inhibit the G1 to S cell cycle phase transition, hindering DNA replication and cell division. DNA viruses, such as adenoviruses, human papillomaviruses, and human cytomegalovirus (HCMV), produce proteins containing the

LxCxE motif that can either degrade the RB protein or inhibit its function, which will help the virus benefit from the host's replication enzymes to replicate its genome [63-65].

4.2.2 G3BP Protein Binding Motif

The Ras GTPase activating protein SH3 domain-binding proteins, known as G3BP, are important for viral replication. The G3BP proteins form a complex and bind to RNA when cells are under environmental stress or viral attack. Upon binding of G3BP to RNA, stress granules are formed to help the cell eliminate the virus and control the viral infection [66]. According to the ELM database, the G3BP binding motif has the pattern [FYLIMV].FG[DES]F [27], often simplified to FGDF. Human herpesvirus [67], Sindbis virus [68,69], Semliki Forest virus [70], and Chikungunya virus [70,71] include FGDF motifs capable of interacting with G3BP and altering its function. The G3BP functional alteration is essential for intracellular viral replication and overcoming the cellular antiviral response [66,67,69,70].

Chikungunya virus, an arbovirus that needs a mosquito vector to be transmitted to a vertebrate host, has two important FGDF motifs in the hypervariable region located towards the C-terminus of nsp3 protein. It has been shown that one FGDF motif is enough to infect the mosquito, but two FGDF motifs are necessary for the virus to be transmitted from mosquito saliva to the vertebrate host [72]. In a relative of Chikungunya virus, Semliki Forest virus, the C-terminal FGDF motif in nsp3 protein is also found to be essential for the interaction with G3BP protein, and without this motif the interaction between G3BP and the replication complex is inhibited [70].

The multiple sequence alignment example shows a variation in the number of FGDF motifs among alphaviruses related to Chikungunya (Figure 4). Further, disorder and secondary structure is not conserved in this hypervariable region suggesting that functional divergence is likely for these FGDF motifs. For instance, in the Chikungunya virus the first motif is found to be in a completely disordered region and the second motif is lacking disorder in only one amino acid based on IUPRED2A predictions with a 0.4 cutoff. However, in the Semliki Forest virus, the two motifs were found to be in ordered protein regions.

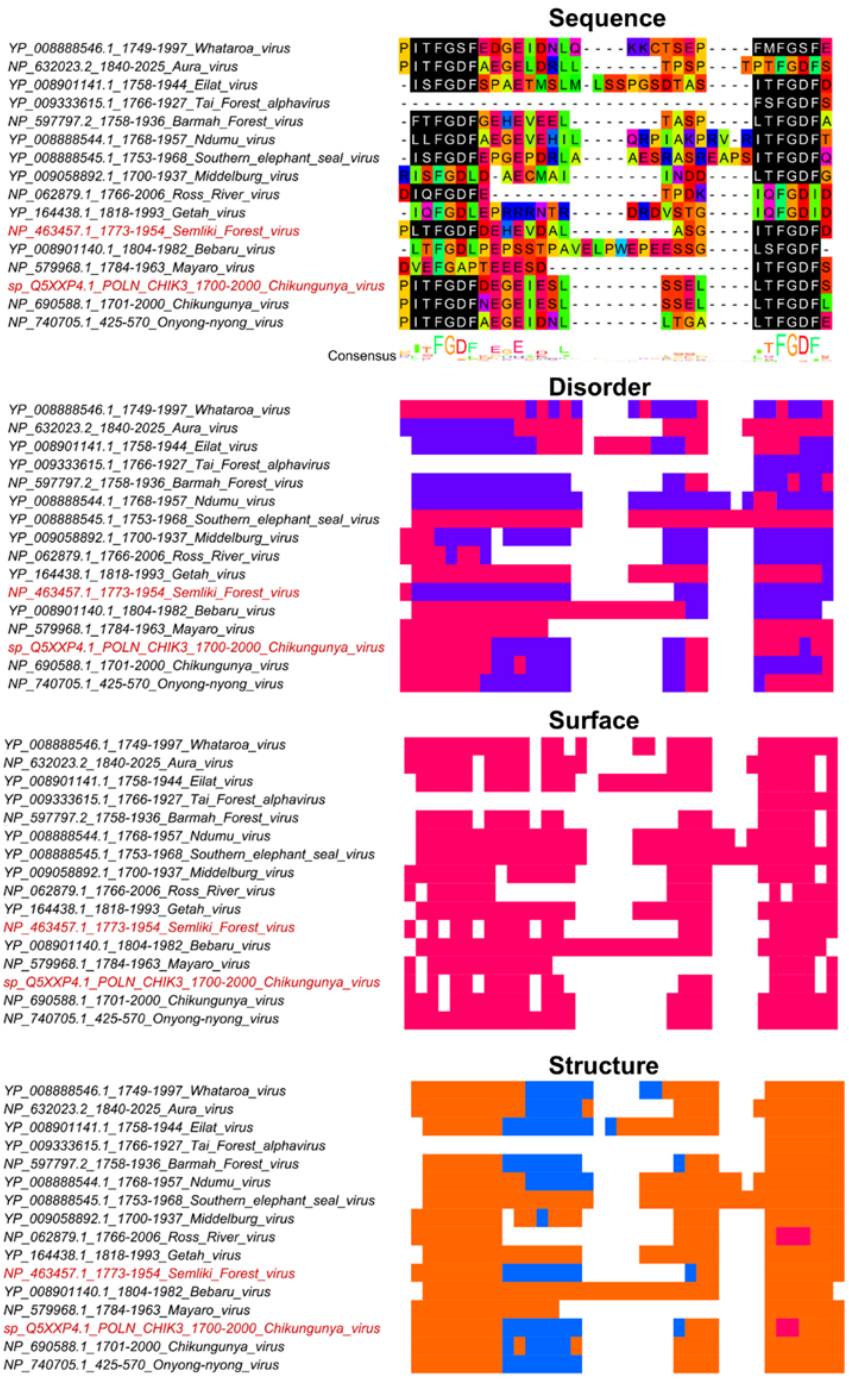

Figure 4. The G3BP binding motif has been verified in the nsp3 protein from Chikungunya virus and Semliki Forest virus from Alphaviruses. Sequences were identified with BLAST using residues 1700–2000 from nsp3 (accession: Q5XXP4) from Chikungunya virus as query. Sequence names shown in red represents true positive instances from the ELM database [27]. The multiple sequence alignment was built with MAFFT+L-INS-i [57] in Jalview [58]. The regular expression pattern

[FYLIMV].FG[DES]F from motif LIG_G3BP_FGDF_1 in the ELM database [27] was identified using Find in Jalview, shown in black with white text. The region shown under Sequence shows the amino acids that corresponds to the true positive motifs from Chikungunya virus and Semliki Forest virus, the connecting amino acids, plus one additional site on each side. The MSA and heatmaps for Disorder, Surface, and Structure are colored as in Figure 3.

In SARS-CoV-2, several studies have reported the interaction of nucleocapsid with the host G3BP proteins [73]. Upon interaction, attenuation of the host immune response occurs due to alteration of the process of stress granules inside the infected cells [74–76]. Kruse et al. proposed that the nucleocapsid-induced inhibition of stress granules is due to the presence of the ΦxFG pattern motif in nucleocapsid, where Φ means any hydrophobic residue, X means any amino acid and the last two amino acids in the motif are phenylalanine and glycine [77]. The motif in SARS-CoV-2 does not follow the last part of the pattern in the ELM database [DES]F, which suggests that the exact pattern is not essential for the functionality of the motif.

4.3. SLiMs and Immune Cell Modulation

Viruses utilize diverse approaches to evade host immunity [78]. One strategy is the use of the pLxIS pattern by Rotaviruses [79]. In humans, the pLxIS motif is found in the stimulator of interferon genes (STING), mitochondrial antiviral signaling protein (MAVS), TIR domain-containing adaptor inducing IFN-β (TRIF), and in interferon regulatory factor 3 (IFN-3). Following the phosphorylation of the pLxIS motif in the adaptor proteins STING, MAVS, and TRIF, they interact with IFN-3 and stimulate the pLxIS motif's phosphorylation in the transcription factor IFN-3. Next, detachment of the adaptor proteins occurs from the IFN-3 protein, followed by IFN-3 homodimerization and activation. Subsequently, the activated IFN-3 dimer transfers to the nucleus and

activates the IFN-β gene's transcription, triggering the release of INF-β from the infected cell and activating the innate immune response [79–81]. In Rotavirus, the pLxIS pattern is observed in the non-structural protein 1 (nsp1) and has the same affinity to IFN-3 as the adaptor proteins; however, when Rotavirus nsp1 pLxIS motif (Figure 5) binds to the IFN-3 protein, ubiquitination and degradation of IFN-3 are initiated. Hence, hindrance of IFN-β transcription occurs, and the virus can effectively escape host defense mechanisms and deactivate one of the innate immune responses [79,82].

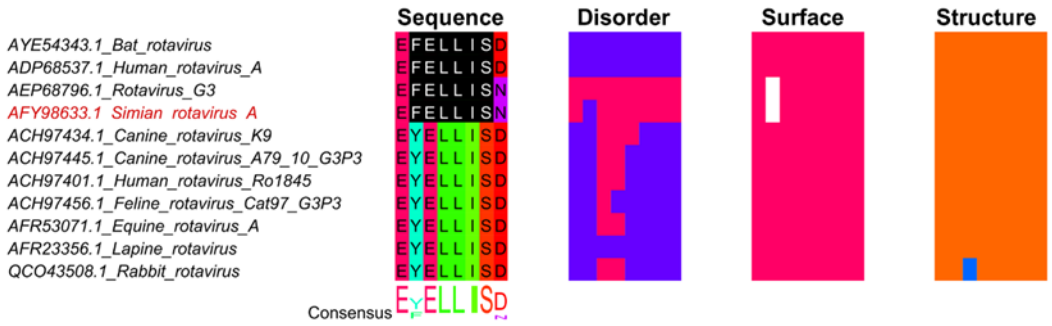

Figure 5. The pLxIS site in nsp1 from Simian rotavirus. Sequences were identified with BLAST using full-length nsp1 from Simian rotavirus (accession: AFY98633.1) as query. Sequence names shown in red represents true positive instances from the ELM database [27]. The multiple sequence alignment was built with MAFFT+L-INS-i [57] in Jalview [58]. The regular expression pattern [VILPF].{1,3}L.I(S) from motif LIG_IRF3_LxIS_1 in the ELM database was identified using Find in Jalview, shown in black with white text. The region shown under Sequence shows the amino acids that corresponds to the true positive motif from Simian rotavirus plus one additional site on each side. The MSA and heatmaps for Disorder, Surface, and Structure are colored as in Figure 3.

4.4. SLiMs Modulating Host Cell Machinery

Although the previous steps are essential in the virus life cycle, viral proteins can also participate in other protein-protein interactions inside the host cell. Viruses can cause unfavorable cellular effects by mediating interactions with other cellular proteins. The following section shows how viruses use different viral-host PPIs to affect the pathogenicity and virulence of a diversity of viruses.

4.4.1 PDZ Binding Motif

PDZ domains are found in a vast number of proteins that recognize a specific C-terminal amino acid pattern [83]. According to the ELM database, the PDZ binding motif pattern is ...[ST].[ACVILF]$ [27]. Proteins that include PDZ domains are involved in numerous cellular processes such as cell signaling pathways, subcellular transport, activating proteases, and recognizing misfolded proteins [83]. Hence, viruses that display a PDZ binding motif (PBM) will have the ability to bind to several PDZ domain containing proteins causing various effects depending on which PDZ domain they interact with [84]. Oncogenic human adenovirus 9 E4 protein and human papillomavirus 18 E6 protein include a PDZ binding motif in their C-terminal regions. Both proteins bind to PDZ domain containing proteins MUPP-1, Dlg, and MAGI-1 [85]. MUPP-1, a multi PDZ domain protein that comprises 13 PDZ domains, is an essential protein for maintaining cell polarity at the tight junction [86]. Dlg, a Drosophila discs large protein and a protein with 3 PDZ domains, is one of the scribble complex proteins, which are involved in maintaining the cellular polarity and adhesion at the cellular junction [87]. MAGI-1 is a membrane associated guanylate cyclase that is located in cellular junction and is important for regulating the proliferation and cellular adhesion between cells [88].

Dlg and MAGI-1 function in tumor suppression [85,87,88]. The binding of human adenovirus 9 to these human proteins inhibits their function through sequestration. Adversely, the E6 protein of some human papillomavirus (HPV) strains that includes the PBM in its C-terminal region will induce these proteins breakdown [85]. Infections with human papillomavirus strains containing PBM in the E6 protein pose a higher risk of causing HPV-associated metastatic cancer. Through the PBM, the E6 protein can perform an interaction with the cellular polarity proteins, leading to loss of cellular polarity and promotion of the proliferation and invasion of cancerous cells [89,90]. The multiple sequence alignment shows that this SLiM is in a highly varying region (Figure 6). The sequence diversity in this region makes it difficult to make a good multiple sequence alignment. Further, intrinsic disorder prediction suggests that this SLiM is not consistently in a disordered region, but the surface accessibility is consistent. Interestingly, the first half of the motif in HPV18 is structured (helix) but the remaining part of the motif is found in a coil state. Such variations may be due to inaccurate predictions but could also be a symptom of functional divergence between the PDZ binding motifs.

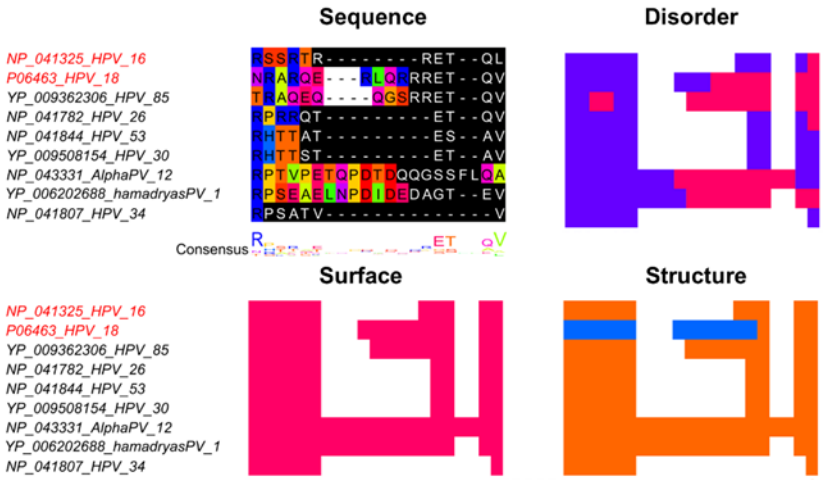

Figure 6. The PDZ domain binding motif in the E6 protein from HPV16 and HPV18. Sequences were identified with BLAST using protein E6 from HPV18 (accession: P06463.1) as query. Sequence names shown in red represents true positive instances from the ELM database [27]. The multiple sequence alignment (MSA) was built with MAFFT+L-INS-i [57] in Jalview [58]. The regular expression pattern …[ST].[ACVILF]$ from motif LIG_PDZ_Class_1 in the ELM database [27] was identified using Find in Jalview, shown in black with white text. The region shown under Sequence shows the amino acids that corresponds to the true positive motif from HPV16 and HPV18 plus one additional site on each side. The MSA and heatmaps for Disorder, Surface, and Structure are colored as in Figure 3.

In SARS-CoV, the envelope protein was found to include PBM, which has the ability to interact with the PDZ domain in the syntenin protein. The interaction of SARS-CoV envelope protein with syntenin was correlated with the P38 MAPK activation, inducing the production of inflammatory cytokines. Mutant PBM motif was correlated with decreased inflammatory response in SARS-CoV infected mice [91]. However, other studies showed that the PBM found in both SARS-CoV and SARS-CoV-2 envelope proteins is capable of interacting with PALS1 protein which is important for maintaining cellular polarity at the cell junction [92–94]. The PBM motif in the envelope protein from SARS-CoV and SARS-CoV-2 has the sequence DLLV [94], which resembles the LIG_PDZ_Class_2 pattern in the ELM database (...[VLIFY].[ACVILF]$) [27], and was found to be in a structurally flexible region that resembles the C-terminal unstructured region in Crumbs protein (Crb-CT). Crb and PATJ protein (PALS1-associated tight junction) binds to PALS1 to form the Crumbs Cell Polarity Complex Component, which is responsible for maintaining cell polarity at the cellular junction [94]. Both the C-terminal BPM motif and the Crb-CT region of the envelope protein were found to bind to PALS1 in a similar fashion [94]. However, the interaction between the envelope protein and PALS1 is thought to cause alteration in the subcellular location of PALS1. The re-

localization of the PALS1 protein to where virus is assembled impedes the cellular junction protein complex formation in the infected epithelial cells. Thus, the infected cell will lose its polarity which can facilitate the viral release from the cells [92–94].

4.4.2 The 14-3-3 Domain-Binding Motif

Another common viral-host interaction is mediated through Serine and Threonine (ST) rich motifs in the 14-3-3 protein family. 14-3-3 proteins are involved in a myriad of signaling pathways and interact with numerous cellular proteins [95–97]. The interaction of the 14-3-3 protein depends on the phosphorylation state of the binding motif. Thus, kinases and phosphatases can affect the motif's binding to the 14-3-3 protein [98]. Regardless of the phosphorylation state of the binding motif, the 14-3-3 SLiM's binding to the 14-3-3 proteins can 1) induce structural changes, 2) block the active site, 3) facilitate the interaction between the motif-containing protein and other proteins, or 4) alter the cellular location of the binding partner [97,98].

In Hepatitis C virus (HCV), the HCV core protein interaction to 14-3-3 protein activates the kinase Raf-1, which induces cellular proliferation and abnormal growth [99]. The HCV genotype 1b core protein has been reported to interact with Raf-1 kinase using the sequence motif RKTpSER, and the phosphorylation of the serine residue was found to be essential for the motif activity [99]. This sequence motif partially overlaps with the R[^DE] {0,2} [^DEPG] ([ST]) (([FWYLMV].) | ([^PRIKGN]P) | ([^PRIKGN].{2,4}[VILMFWYP])) pattern of the canonical 14-3-3 binding motif (LIG_14-3-3_CanoR_1) in the ELM database [27].

In SARS-CoV, binding of the 14-3-3 domain-containing proteins to the phosphorylated nucleocapsid is involved in translocation of the nucleocapsid protein

between the cytoplasm and nucleus, altering the functionality of the 14-3-3 interacting protein [100]. In the closely related SARS-CoV-2, nucleocapsid has not yet been detected in the nucleus, but it has been found to interact with various 14-3-3 protein isoforms in the cytoplasm [101]. Several sequence patterns identified in both SARS-CoV and SARS-CoV-2 are found in a disordered S/T-rich protein region of the nucleocapsid protein with multiple known phosphorylation sites [101] and resemble, either partially or completely, the canonical 14-3-3 pattern found in the ELM database [27]. The phosphorylation and disorder property of the presented motifs suggest a similarity to other 14-3-3 binding motifs where phosphorylation and disorder are essential for interacting with the 14-3-3 domain-containing proteins [102]. Although no viral 14-3-3 binding motif examples are included yet in the ELM database, these examples highlight that viruses may have numerous molecular effects on cells through interactions with 14-3-3, mediated by SLiMs.

4.5. SLiMs Responsible for Viral Exit from the Cell

Viruses have several strategies to egress their host cells, which can be achieved through cell lysis, budding from the cell membrane, or exocytosis using the secretory pathway. SLiMs can enhance viral egress through budding. One example is the interaction of the viral proteins with the endosomal sorting complexes required for transport (ESCRT) pathway inside the cell. The importance of viral late domains (L domains) has been widely implicated in the viral budding process, and short sequence motifs, P[TS]AP, PPxY, and LYPxL, have been involved in the interaction with the ESCRT pathway machinery [103,104,105]. Such motifs were found to be highly conserved across diverse types of viruses, including Poxviruses [106], Hepatitis C viruses

[107], Rhabdoviruses [108], Retroviruses [109], Arenaviruses [110], and Filoviruses [111,112]. Ebola VP40 (Figure 7) and HIV-1 contain PPxY motifs that are recognized by a highly conserved enzyme in humans (E3 ubiquitin ligase) [113,114]. E3 ubiquitin ligase enzyme is involved in regulating a plethora of biological processes by stimulating the ubiquitination and subsequent degradation of their target protein [115]. Interactions with the WW domain of ubiquitin ligase enzymes, recruitment of Tsg101, and the ubiquitination by specific ubiquitin ligase enzymes have been shown to facilitate the ESCRT pathway-mediated viral budding [113,114]. The role of ESCRT pathway and viral late domains in viral exit have been extensively reviewed [103,116], including the importance of the ESCRT pathway in different phases of the viral life cycle [117].

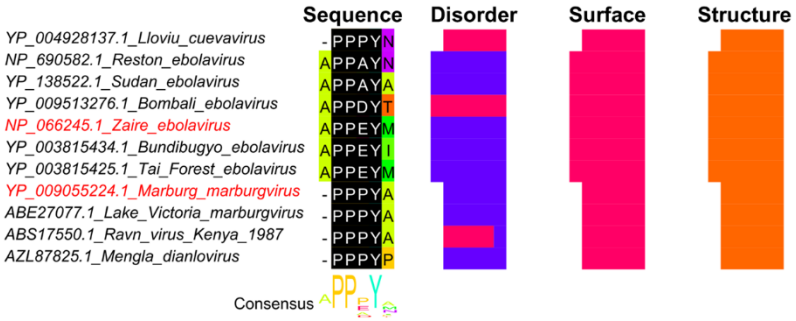

Figure 7. The PPxY motif in the matrix protein VP40 from Ebola virus. Sequences were identified with BLAST using full-length VP40 from Ebola virus (accession: Q05128) as query against the refseq_protein and nr databases. Sequence names shown in red represents true positive instances from the ELM database [27]. The multiple sequence alignment was built with MAFFT+L-INS-i [57] in Jalview [58]. The regular expression pattern PP.Y from motif LIG_WW_1 in the ELM database [27] was identified using Find in Jalview, shown in black with white text. The region shown under Sequence corresponds to the true positive motif from Zaire Ebola virus and Marburg marburg virus plus one additional site on each side. It should be noted that query protein Q05128 Uniprot ID is identical to protein NP_066245.1 used in the multiple sequence alignment.

5. Conclusion and Future Perspective

The small genome size of viruses and their inability to replicate outside a host go hand in hand with their need to hijack host cell machinery [2]. SLiMs with varying evolutionary rates in different viral families can mutate to accommodate various selective pressures stemming from their environment. The fitness of viruses depends on their capacity to alter host cell machinery and escape detection by the immune system. This capacity is governed, in part, by the potential to mimic and compete with functionally important protein interactions. In this review, we highlighted the importance of viral mimicry mediated by SLiMs at select steps of the virus life cycle. We also showed how specific SLiMs might affect virulence and pathogenicity. These SLiM actions are mediated by viral-host protein-protein interactions.

Previous studies on eukaryotic SLiMs showed that physicochemical properties, such as secondary structure and disorder, should be considered when studying SLiMs as the majority of the functionally verified SLiMs were found to be disordered and enriched with polar residues [34]. Based on disorder predictions, the true positive experimentally verified viral SLiMs deposited in the ELM database are not necessarily intrinsically disordered, but they are surface exposed and mainly in a conformationally flexible coil rather than in alpha helices or beta strands. Our findings for the viral SLiMs give rise to questions regarding disorder content and other structural characteristics of the corresponding eukaryotic linear motifs in the hosts of viruses, and for eukaryotic linear motifs, in general. The ELM database has grown rapidly over the last 10 years and re-analysis of disorder content is warranted. Among the viral SLiMs, the most abundant categories are the ligand binding sites and post-translationally modified sites. Ligand

binding sites are the most common class among the fully disordered sites, while the post-translational modification sites are the most common among the fully ordered sites. Given that disorder content appears to vary between different functional classes of motifs, an analysis into disorder content variation across these classes may illuminate function-specific traits of importance in differentiating false and true positive SLiMs.

Proteomes from eukaryotes tend to have more disorder content overall than proteomes from bacteria and viruses [118,119]. It is possible that the disorder content required for SLiMs to be functional not only depends on the identity of the SLiM, but also on other contexts such as genome complexity and overall disorder content of the proteome. Eukaryotic genomes evolve under multifaceted constraint that differ from the constraint acting on viruses [120]. For eukaryotes, disordered regions are often able to participate in multiple distinct PPIs [121]. Disorder is advantageous at binding interfaces that rely on conformational transitions where SLiMs controlled by post-translational modifications may act as molecular on/off switches [122]. However, disorder may become less advantageous when an ordered viral SLiM mimics a functional conformation of a host SLiM so that it is always switched on or off.

We showed an example of the G3BP binding motif in Chikungunya virus and Semliki Forest virus. Based on IUPRED2A, the former is found in a disordered region, while the latter is in an ordered region (Figure 4). Since intrinsic disorder is not conserved, changes in disorder can potentially change the functional potential of a SLiM; however, intrinsic disorder may not be as important for viral SLiMs as often stated. The majority of the experimentally confirmed viral SLiMs were almost entirely found in a surface accessible coil region, unlike disorder where at least 1 in 4 motifs was devoid of disorder. HIV-1

envelope furin cleavage site motif and E6 HPV 18 PBM were predicted to be a mix of both coil and helix, which poses a question about the differences between flexible and disordered protein regions, and whether flexibility and disorder should both be considered when searching for functional SLiMs.

Experimental verification of viral SLiMs can be challenging. The large SARS-CoV-2 dataset that has accumulated since this virus emerged in late 2019 has a wealth of information. Currently, some SLiMs for SARS-CoV-2 have been verified [123,124]. We expect that more are to come and that they will contribute to how we analyze viral SLiMs. For example, the subcellular location of most SARS-CoV-2 proteins have been determined (Figure 8). The Cell Atlas [125] and the Human Protein Atlas [126] provide subcellular locations and more for human proteins. Combining information about shared cellular locations will further illuminate potential viral network interference in the host cell. Computational methods provide time- and cost-effective, low-risk ways to predict the presence and function of these crucial motifs, which may be experimentally verified in vitro.

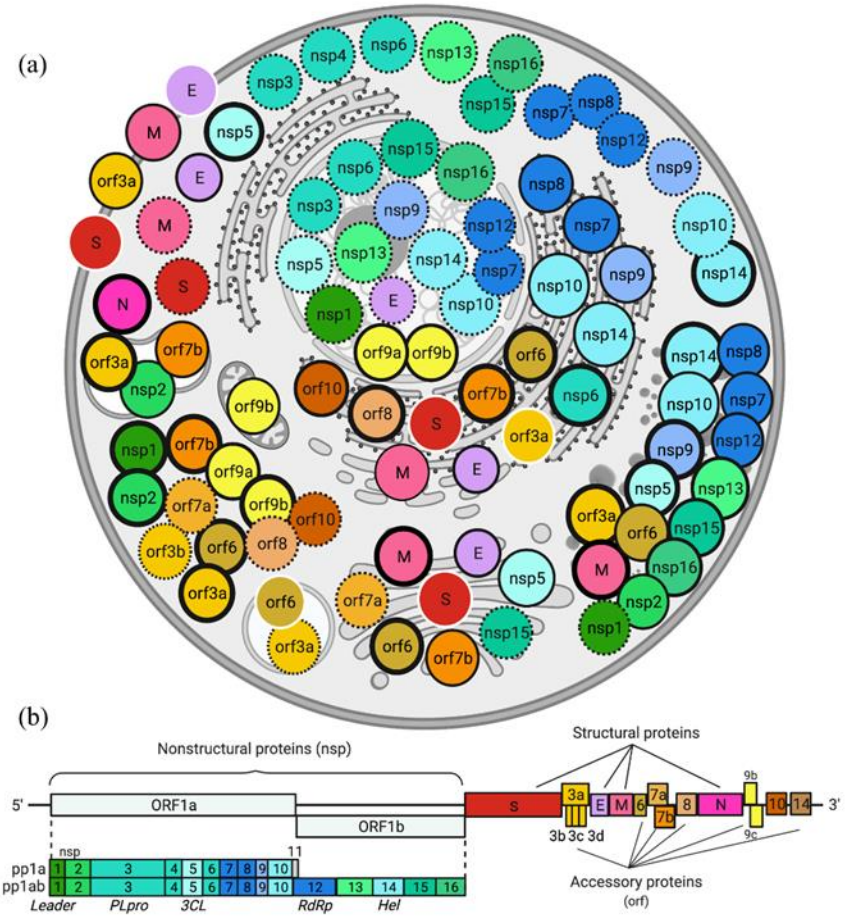

Figure 8. Cellular context. Subcellular localization of SARS-CoV-2 proteins (circles) in human cells based on experimental data (thick border: multiple sources; dotted border: [127]; thin black border: [128]; white border: [129,130,131]). (a). Each protein is colored as in the SARS-CoV-2 proteome (b). Proteins that form complexes are colored similarly; nsp 3/4/6, nsp 7/8/12, nsp 10/14. SARS-CoV-2 proteins localize to the following organelles: lysosome (nsp2, orf3a, and orf7b), endosome (orf3a and orf6), plasma membrane (envelope (E), membrane (M), spike (S), and orf3a), Golgi apparatus (E, M, S, nsp5, nsp15, orf6, orf7a, and orf7b), endoplasmic reticulum (E, M, S, nsp6-10, nsp14, orf6, orf7b, orf8, and orf10), nucleolus (E, nsp1, nsp3, nsp5-7, nsp9-10, nsp12-16 and orf9a-9b), punctate cytoplasm (M, nsp1, nsp2, nsp5, nsp7-10, nsp12-16, orf3a, and orf6), and diffuse cytoplasm (E, M, nucleocapsid (N), S, nsp1-16, nsp10, nsp12-16, orf3a-3b, orf6, orf7a-7b, orf8, orf9a-9b, and orf10). Created with BioRender.com (accessed on 30 October 2021).

While the limitations of both computational and experimental approaches of linear motifs must be closely considered to decrease the probability of misleading false positive results, predictions of SLiMs have proven helpful in elucidating how SARS-

CoV-2 interacts with its human host (e.g., [44,61,128,132]). Altogether, this review shows the promise for how molecular mimicry discovery in different viral families can improve our understanding of the virus-host interface.

Supplementary Materials: The following are available online at

https://www.mdpi.com/article/10.3390/v13122369/s1, Table S1:

viral_slim_predictions.csv

REFERENCES

1. Ryu, W.-S. Virus Life Cycle. *Mol. Virol. Hum. Pathog. Viruses* **2017**, 31, doi:10.1016/B978-0-12-800838-6.00003-5.

2. Davey, N.E.; Travé, G.; Gibson, T.J. How viruses hijack cell regulation. *Trends Biochem. Sci.* **2011**, *36*, 159–169.

3. Cui, J.; Schlub, T.E.; Holmes, E.C. An Allometric Relationship between the Genome Length and Virion Volume of Viruses. *J. Virol.* **2014**, *88*, 6403, doi:10.1128/JVI.00362-14.

4. Xue, B.; Blocquel, D.; Habchi, J.; Uversky, A. V; Kurgan, L.; Uversky, V.N.; Longhi, S. Structural disorder in viral proteins. *Chem. Rev.* **2014**, *114*, 6880–6911, doi:10.1021/cr4005692.

5. Gago, S.; Elena, S.F.; Flores, R.; Sanjuán, R. Extremely high mutation rate of a hammerhead viroid. *Science.* **2009**, *323*, 1308, doi:10.1126/science.1169202.

6. Sanjuán, R.; Domingo-Calap, P. Mechanisms of viral mutation. *Cell. Mol. Life Sci.* **2016**, *73*, 4433, doi:10.1007/S00018-016-2299-6.

7. Pfeiffer, J.K.; Kirkegaard, K. Increased Fidelity Reduces Poliovirus Fitness and Virulence under Selective Pressure in Mice. *PLOS Pathog.* **2005**, *1*, e11, doi:10.1371/JOURNAL.PPAT.0010011.

8. Furió, V.; Moya, A.; Sanjuán, R. The cost of replication fidelity in an RNA virus. *Proc. Natl. Acad. Sci.* **2005**, *102*, 10233–10237, doi:10.1073/PNAS.0501062102.

9. Guven-Maiorov, E.; Tsai, C.-J.; Nussinov, R. Structural host-microbiota interaction networks. **2017**, doi:10.1371/journal.pcbi.1005579.

10. Franzosa, E.A.; Xia, Y. Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10538–10543, doi:10.1073/pnas.1101440108.

11. Hagai, T.; Azia, A.; Babu, M.M.; Andino, R. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell Rep.* **2014**, *7*, 1729–1739, doi:10.1016/j.celrep.2014.04.052.

12. Hraber, P.; O'Maille, P.E.; Silberfarb, A.; Davis-Anderson, K.; Generous, N.; McMahon, B.H.; Fair, J.M. Resources to Discover and Use Short Linear Motifs in Viral Proteins. *Trends Biotechnol.* **2020**, *38*, 113–127.

13. Sobhy, H. A review of functional motifs utilized by viruses. *Proteomes* **2016**, *4*.

14. Becerra, A.; Bucheli, V.A.; Moreno, P.A. Prediction of virus-host protein-protein interactions mediated by short linear motifs. *BMC Bioinformatics* **2017**, *18*, doi:10.1186/s12859-017-1570-7.

15. Uversky, V.N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Front. Phys.* **2019**, *0*, 10, doi:10.3389/FPHY.2019.00010.

16. Fuxreiter, M.; Tompa, P.; Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **2007**, *23*, 950–956, doi:10.1093/bioinformatics/btm035.

17. Davey, N.E. The functional importance of structure in unstructured protein regions. *Curr. Opin. Struct. Biol.* **2019**, *56*, 155–163, doi:10.1016/J.SBI.2019.03.009.

18. Hsu, W.-L.; Oldfield, C.J.; Xue, B.; Meng, J.; Huang, F.; Romero, P.; Uversky, V.N.; Dunker, A.K. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci.* **2013**, *22*, 258–273, doi:10.1002/pro.2207.

19. Maheshwari, S.; Brylinski, M. Predicting protein interface residues using easily accessible on-line resources. *Brief. Bioinform.* **2015**, *16*, 1025, doi:10.1093/BIB/BBV009.

20. Chemes, L.B.; De Prat-Gay, G.; Sá Nchez, I.E. Convergent evolution and mimicry of protein linear motifs in host-pathogen interactions This review comes from a themed issue on Sequences and topology. *Curr. Opin. Struct. Biol.* **2015**, *32*, 91–101, doi:10.1016/j.sbi.2015.03.004.

21. Pushker, R.; Mooney, C.; Davey, N.E.; Jacqué, J.-M.; Shields, D.C. Marked variability in the extent of protein disorder within and between viral families. *PLoS One* **2013**, *8*, e60724, doi:10.1371/journal.pone.0060724.

22. Rahaman, J.; Siltberg-Liberles, J. Avoiding regions symptomatic of conformational and functional flexibility to identify antiviral targets in current and future coronaviruses. *Genome Biol. Evol.* **2016**, *8*, 3471–3484, doi:10.1093/gbe/evw246.

23. Perkins, J.R.; Diboun, I.; Dessailly, B.H.; Lees, J.G.; Orengo, C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure* **2010**, *18*, 1233–1243, doi:10.1016/J.STR.2010.08.007.

24. Hugo, W.; Sung, W.K.; Ng, S.K. Discovering interacting domains and motifs in protein-protein interactions. *Methods Mol. Biol.* **2013**, *939*, 9–20, doi:10.1007/978-1-62703-107-3_2.

25. Budayeva, H.G.; Cristea, I.M. A mass spectrometry view of stable and transient protein interactions. *Adv. Exp. Med. Biol.* **2014**, *806*, 263, doi:10.1007/978-3-319-06068-2_11.

26. Davey, N.E.; Seo, M.-H.; Yadav, V.K.; Jeon, J.; Nim, S.; Krystkowiak, I.; Blikstad, C.; Dong, D.; Markova, N.; Kim, P.M.; et al. Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. *FEBS J.* **2017**, *284*, 485–498, doi:10.1111/FEBS.13995.

27. Kumar, M.; Gouw, M.; Michael, S.; Sámano-Sánchez, H.; Pancsa, R.; Glavina, J.; Diakogianni, A.; Valverde, J.A.; Bukirova, D.; Signalyševa, J.; et al. ELM-the eukaryotic linear motif resource in **2020**. *Nucleic Acids Res.* 2020, *48*, D296–D306, doi:10.1093/nar/gkz1030.

28. Krystkowiak, I.; Davey, N.E. SLiMSearch: A framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Res.* **2017**, *45*, W464–W469, doi:10.1093/nar/gkx238.

29. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, doi:10.1093/nar/gkp335.

30. Davey, N.E.; Haslam, N.J.; Shields, D.C.; Edwards, R.J. SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic Acids Res.* **2010**, *38*, W534–W539, doi:10.1093/NAR/GKQ440.

31. Edwards, R.J.; Palopoli, N. Computational prediction of Short Linear Motifs from protein sequences. *Methods Mol. Biol.* **2015**, *1268*, 89–141, doi:10.1007/978-1-4939-2285-7_6.

32. Edwards, R.J.; Paulsen, K.; Aguilar Gomez, C.M.; Pérez-Bercoff, Å. Computational Prediction of Disordered Protein Motifs Using SLiMSuite. *Methods Mol. Biol.* **2020**, *2141*, 37–72, doi:10.1007/978-1-0716-0524-0_3.

33. Erdős, G.; Pajkos, M.; Dosztányi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **2021**, *49*, W297–W303, doi:10.1093/NAR/GKAB408.

34. Davey, N.E.; Van Roey, K.; Weatheritt, R.J.; Toedt, G.; Uyar, B.; Altenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T.J. Attributes of short linear motifs. *Mol. Biosyst.* **2012**, *8*, 268–281, doi:10.1039/c1mb05231d.

35. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337, doi:10.1093/nar/gky384.

36. Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **2005**, *347*, 827–839, doi:10.1016/j.jmb.2005.01.071.

37. Klausen, M.S.; Jespersen, M.C.; Nielsen, H.; Jensen, K.K.; Jurtz, V.I.; Sønderby, C.K.; Sommer, M.O.A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinforma.* **2019**, *87*, 520–527, doi:10.1002/prot.25674.

38. Kotecha, A.; Wang, Q.; Dong, X.; Ilca, S.L.; Ondiviela, M.; Zihe, R.; Seago, J.; Charleston, B.; Fry, E.E.; Abrescia, N.G.A.; et al. Rules of engagement between αvβ6 integrin and foot-and-mouth disease virus. *Nat. Commun. 2017 81* **2017**, *8*, 1–8, doi:10.1038/ncomms15408.

39. Chesnokova, L.S.; Nishimura, S.L.; Hutt-Fletcher, L.M. Fusion of epithelial cells by Epstein–Barr virus proteins is triggered by binding of viral glycoproteins gHgL to integrins αvβ6 or αvβ8. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 20464, doi:10.1073/PNAS.0907508106.

40. Hussein, H.A.M.; Walker, L.R.; Abdel-Raouf, U.M.; Desouky, S.A.; Montasser, A.K.M.; Akula, S.M. Beyond RGD: virus interactions with integrins. *Arch. Virol.* **2015**, *160*, 2669, doi:10.1007/S00705-015-2579-8.

41. Barczyk, M.; Carracedo, S.; Gullberg, D. Integrins. *Cell Tissue Res. 2009 3391* **2009**, *339*, 269–280, doi:10.1007/S00441-009-0834-6.

42. Stewart, P.L.; Nemerow, G.R. Cell integrins: commonly used receptors for diverse viral pathogens. *Trends Microbiol.* **2007**, *15*, 500–507, doi:10.1016/J.TIM.2007.10.001.

43. Sigrist, C.J.; Bridge, A.; Le Mercier, P. A potential role for integrins in host cell entry by SARS-CoV-2. *Antiviral Res.* **2020**, *177*, doi:10.1016/j.antiviral.2020.104759.

44. Makowski, L.; Olson-Sidford, W.; Weisel, J.W. Biological and clinical consequences of integrin binding via a rogue rgd motif in the sars cov-2 spike protein. *Viruses* **2021**, *13*.

45. Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nat. 2020 5817807* **2020**, *581*, 215–220, doi:10.1038/s41586-020-2180-5.

46. Thomas, G. Furin at the cutting edge: From protein traffic to embryogenesis and disease. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 753–766.

47. Braun, E.; Sauter, D. Furin-mediated protein processing in infectious diseases and cancer. *Clin. Transl. Immunol.* **2019**, *8*.

48. Moulard, M.; Decroly, E. Maturation of HIV envelope glycoprotein precursors by cellular endoproteases. *Biochim. Biophys. Acta - Rev. Biomembr.* **2000**, *1469*, 121–132, doi:10.1016/S0304-4157(00)00014-9.

49. Cavanagh, D.; Davis, P.J.; Pappin, D.J.C.; Binns, M.M.; Boursnell, M.E.G.; Brown, T.D.K. Coronavirus IBV: Partial amino terminal sequencing of spike polypeptide S2 identifies the sequence Arg-Arg-Phe-Arg-Arg at the cleavage site of the spike precursor propolypeptide of IBV strains Beaudette and M41. *Virus Res.* **1986**, *4*, 133–143, doi:10.1016/0168-1702(86)90037-7.

50. Zybert, I.A.; Ende-Metselaar, H. van der; Wilschut, J.; Smit, J.M. Functional importance of dengue virus maturation: infectious properties of immature virions. *J. Gen. Virol.* **2008**, *89*, 3047–3051, doi:10.1099/VIR.0.2008/002535-0.

51. Gordon, V.M.; Leppla, S.H. Proteolytic activation of bacterial toxins: Role of bacterial and host cell proteases. *Infect. Immun.* **1994**, *62*, 333–340.

52. Tsuneoka, M.; Nakayama, K.; Hatsuzawa, K.; Komada, M.; Kitamura, N.; Mekada, E. Evidence for involvement of furin in cleavage and activation of diphtheria toxin. *J. Biol. Chem.* **1993**, *268*, 26461–26465, doi:10.1016/S0021-9258(19)74337-3.

53. Izaguirre, G. The Proteolytic Regulation of Virus Cell Entry by Furin and Other Proprotein Convertases. *Viruses* **2019**, *11*, doi:10.3390/V11090837.

54. Tian, S. A 20 Residues Motif Delineates the Furin Cleavage Site and its Physical Properties May Influence Viral Fusion: *Biochem. Insights* **2009**, *2*, doi:10.4137/BCI.S2049.

55. Mukherjee, S.; Sirohi, D.; Dowd, K.A.; Chen, Z.; Diamond, M.S.; Kuhn, R.J.; Pierson, T.C. Enhancing dengue virus maturation using a stable furin over-expressing cell line. *Virology* **2016**, *497*, 33–40, doi:10.1016/j.virol.2016.06.022.

56. Tse, L. V.; Hamilton, A.M.; Friling, T.; Whittaker, G.R. A Novel Activation Mechanism of Avian Influenza Virus H9N2 by Furin. *J. Virol.* **2014**, *88*, 1673–1683, doi:10.1128/jvi.02648-13.

57. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–80, doi:10.1093/molbev/mst010.

58. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*, 1189–1191, doi:10.1093/bioinformatics/btp033.

59. Nunez-Castilla, J.; Siltberg-Liberles, J. An easy protocol for evolutionary analysis of intrinsically disordered proteins. In *Methods in Molecular Biology*; Humana Press Inc., **2020**; Vol. 2141, pp. 147–177.

60. Coutard, B.; Valle, C.; de Lamballerie, X.; Canard, B.; Seidah, N.G.; Decroly, E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* **2020**, *176*, doi:10.1016/J.ANTIVIRAL.2020.104742.

61. Papa, G.; Mallery, D.L.; Albecka, A.; Welch, L.G.; Cattin-Ortolá, J.; Luptak, J.; Paul, D.; McMahon, H.T.; Goodfellow, I.G.; Carter, A.; et al. Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion. *PLOS Pathog.* **2021**, *17*, e1009246, doi:10.1371/JOURNAL.PPAT.1009246.

62. Wu, Y.; Zhao, S. Furin cleavage sites naturally occur in coronaviruses. *Stem Cell Res.* **2021**, *50*, 102115, doi:10.1016/J.SCR.2020.102115.

63. Felsani, A.; Mileo, A.M.; Paggi, M.G. Retinoblastoma family proteins as key targets of the small DNA virus oncoproteins. *Oncogene* **2006**, *25*, 5277–5285.

64. Lee, J.O.; Russo, A.A.; Pavletich, N.P. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature* **1998**, *391*, 859–865, doi:10.1038/36038.

65. VanDeusen, H.R.; Kalejta, R.F. The Retinoblastoma Tumor Suppressor Promotes Efficient Human Cytomegalovirus Lytic Replication. *J. Virol.* **2015**, *89*, 5012–5021, doi:10.1128/jvi.00175-15.

66. Panas, M.D.; Schulte, T.; Thaa, B.; Sandalova, T.; Kedersha, N.; Achour, A.; McInerney, G.M. Viral and Cellular Proteins Containing FGDF Motifs Bind G3BP to Block Stress Granule Formation. *PLoS Pathog.* **2015**, *11*, doi:10.1371/journal.ppat.1004659.

67. Finnen, R.L.; Pangka, K.R.; Banfield, B.W. Herpes Simplex Virus 2 Infection Impacts Stress Granule Accumulation. *J. Virol.* **2012**, *86*, 8119, doi:10.1128/JVI.00313-12.

68. Cristea, I.M.; Carroll, J.-W.N.; Rout, M.P.; Rice, C.M.; Chait, B.T.; MacDonald, M.R. Tracking and Elucidating Alphavirus-Host Protein Interactions. *J. Biol. Chem.* **2006**, *281*, 30269–30278, doi:10.1074/JBC.M603980200.

69. Panas, M.D.; Ahola, T.; McInerney, G.M. The C-Terminal Repeat Domains of nsP3 from the Old World Alphaviruses Bind Directly to G3BP. *J. Virol.* **2014**, *88*, 5888, doi:10.1128/JVI.00439-14.

70. Panas, M.D.; Varjak, M.; Lulla, A.; Eng, K.E.; Merits, A.; Hedestam, G.B.K.; McInerney, G.M. Sequestration of G3BP coupled with efficient translation inhibits stress granules in Semliki Forest virus infection. *Mol. Biol. Cell* **2012**, *23*, 4701–4712, doi:10.1091/mbc.E12-08-0619.

71. Fros, J.J.; Domeradzka, N.E.; Baggen, J.; Geertsema, C.; Flipse, J.; Vlak, J.M.; Pijlman, G.P. Chikungunya Virus nsP3 Blocks Stress Granule Assembly by Recruitment of G3BP into Cytoplasmic Foci. *J. Virol.* **2012**, *86*, 10873–10879, doi:10.1128/jvi.01506-12.

72. Göertz, G.P.; Lingemann, M.; Geertsema, C.; Abma-Henkens, M.H.C.; Vogels, C.B.F.; Koenraadt, C.J.M.; Oers, M.M. van; Pijlman, G.P. Conserved motifs in the hypervariable domain of chikungunya virus nsP3 required for transmission by Aedes aegypti mosquitoes. *PLoS Negl. Trop. Dis.* **2018**, *12*, e0006958, doi:10.1371/JOURNAL.PNTD.0006958.

73. Gordon, D.E.; Jang, G.M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K.M.; O'Meara, M.J.; Rezelj, V. V.; Guo, J.Z.; Swaney, D.L.; et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **2020**, *583*, 459–468, doi:10.1038/s41586-020-2286-9.

74. Cai, T.; Yu, Z.; Wang, Z.; Liang, C.; Richard, S. Arginine methylation of SARS-Cov-2 nucleocapsid protein regulates RNA binding, its ability to suppress stress granule formation, and viral replication. *J. Biol. Chem.* **2021**, *297*, doi:10.1016/J.JBC.2021.100821.

75. Kamel, W.; Noerenberg, M.; Cerikan, B.; Chen, H.; Järvelin, A.I.; Kammoun, M.; Lee, J.Y.; Shuai, N.; Garcia-Moreno, M.; Andrejeva, A.; et al. Global analysis of protein-RNA interactions in SARS-CoV-2-infected cells reveals key regulators of infection. *Mol. Cell* **2021**, *81*, 2851-2867.e7, doi:10.1016/J.MOLCEL.2021.05.023.

76. Zheng, X.; Sun, Z.; Yu, L.; Shi, D.; Zhu, M.; Yao, H.; Li, L. Interactome Analysis of the Nucleocapsid Protein of SARS-CoV-2 Virus. **2021**, *10*, 1155, doi:10.3390/PATHOGENS10091155.

77. Kruse, T.; Benz, C.; Garvanska, D.H.; Lindqvist, R.; Mihalic, F.; Coscia, F.; Inturi, R.T.; Sayadi, A.; Simonetti, L.; Nilsson, E.; et al. Large scale discovery of coronavirus-host factor protein interaction motifs reveals SARS-CoV-2 specific mechanisms and vulnerabilities. *bioRxiv* **2021**, 2021.04.19.440086, doi:10.1101/2021.04.19.440086.

78. Alcami, A.; Koszinowski, U.H.; Alcami, A.; Koszinowski, U.H. Viral mechanisms of immune evasion. *Trends Microbiol.* **2000**, *8*, 410–418, doi:10.1016/S0966-842X(00)01830-8.

79. Zhao, B.; Shu, C.; Gao, X.; Sankaran, B.; Du, F.; Shelton, C.L.; Herr, A.B.; Ji, J.-Y.; Li, P. Structural basis for concerted recruitment and activation of IRF-3 by innate immune adaptor proteins. *Proc. Natl. Acad. Sci.* **2016**, *113*, E3403–E3412, doi:10.1073/PNAS.1603269113.

80. Rackov, G.; Shokri, R.; De Mon, M.Á.; Martínez-A., C.; Balomenos, D. The Role of IFN-β during the Course of Sepsis Progression and Its Therapeutic Potential. *Front. Immunol.* **2017**, *8*, 493, doi:10.3389/FIMMU.2017.00493.

81. Liu, S.; Cai, X.; Wu, J.; Cong, Q.; Chen, X.; Li, T.; Du, F.; Ren, J.; Wu, Y.T.; Grishin, N. V.; et al. Phosphorylation of innate immune adaptor proteins MAVS, STING, and TRIF induces IRF3 activation. *Science.* **2015**, *347*, doi:10.1126/science.aaa2630.

82. Barro, M.; Patton, J.T. Rotavirus nonstructural protein 1 subverts innate immune response by inducing degradation of IFN regulatory factor 3. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 4114–4119, doi:10.1073/pnas.0408376102.

83. Lee, H.-J.; Zheng, J.J. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Commun. Signal.* **2010**, *8*, 1–18, doi:10.1186/1478-811X-8-8.

84. Charbonnier, S.; Nominé, Y.; Ramírez, J.; Luck, K.; Chapelle, A.; Stote, R.H.; Travé, G.; Kieffer, B.; Atkinson, R.A. The Structural and Dynamic Response of MAGI-1 PDZ1 with Noncanonical Domain Boundaries to the Binding of Human Papillomavirus E6. *J. Mol. Biol.* **2011**, *406*, 745–763, doi:10.1016/J.JMB.2011.01.015.

85. Lee, S.S.; Glaunsinger, B.; Mantovani, F.; Banks, L.; Javier, R.T. Multi-PDZ Domain Protein MUPP1 Is a Cellular Target for both Adenovirus E4-ORF1 and High-Risk Papillomavirus Type 18 E6 Oncoproteins. *J. Virol.* **2000**, *74*, 9680, doi:10.1128/JVI.74.20.9680-9693.2000.

86. Hamazaki, Y.; Itoh, M.; Sasaki, H.; Furuse, M.; Tsukita, S. Multi-PDZ Domain Protein 1 (MUPP1) Is Concentrated at Tight Junctions through Its Possible Interaction with Claudin-1 and Junctional Adhesion Molecule. *J. Biol. Chem.* **2002**, *277*, 455–461, doi:10.1074/JBC.M109005200.

87. Su, W.-H.; Mruk, D.D.; Wong, E.W.P.; Lui, W.-Y.; Cheng, C.Y. Polarity Protein Complex Scribble/Lgl/Dlg and Epithelial Cell Barriers. *Adv. Exp. Med. Biol.* **2012**, *763*, 149.

88. Wörthmüller, J.; Rüegg, C. MAGI1, a Scaffold Protein with Tumor Suppressive and Vascular Functions. *Cells* **2021**, *10*, 1494, doi:10.3390/CELLS10061494.

89. Ganti, K.; Broniarczyk, J.; Manoubi, W.; Massimi, P.; Mittal, S.; Pim, D.; Szalmas, A.; Thatte, J.; Thomas, M.; Tomaić, V.; et al. The human papillomavirus E6 PDZ binding motif: From life cycle to malignancy. *Viruses* **2015**, *7*, 3530–3551, doi:10.3390/v7072785.

90. Narisawa-Saito, M.; Kiyono, T. Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: Roles of E6 and E7 proteins. *Cancer Sci.* **2007**, *98*, 1505–1511.

91. Jimenez-Guardeño, J.M.; Nieto-Torres, J.L.; DeDiego, M.L.; Regla-Nava, J.A.; Fernandez-Delgado, R.; Castaño-Rodriguez, C.; Enjuanes, L. The PDZ-Binding Motif of Severe Acute Respiratory Syndrome Coronavirus Envelope Protein Is a Determinant of Viral Pathogenesis. *PLoS Pathog.* **2014**, *10*, 1004320, doi:10.1371/JOURNAL.PPAT.1004320.

92. Javorsky, A.; Humbert, P.O.; Kvansakul, M. Structural basis of coronavirus E protein interactions with human PALS1 PDZ domain. *Commun. Biol. 2021 41* **2021**, *4*, 1–8, doi:10.1038/s42003-021-02250-7.

93. Teoh, K.-T.; Siu, Y.-L.; Chan, W.-L.; Schlüter, M.A.; Liu, C.-J.; Peiris, J.S.M.; Bruzzone, R.; Margolis, B.; Nal, B. The SARS Coronavirus E Protein Interacts with PALS1 and Alters Tight Junction Formation and Epithelial Morphogenesis. *Mol. Biol. Cell* **2010**, *21*, 3838–3852, doi:10.1091/MBC.E10-04-0338.

94. Chai, J.; Cai, Y.; Pang, C.; Wang, L.; McSweeney, S.; Shanklin, J.; Liu, Q. Structural basis for SARS-CoV-2 envelope protein recognition of human cell junction protein PALS1. *Nat. Commun.* **2021**, *12*, 1–6, doi:10.1038/s41467-021-23533-x.

95. Fu, H.; Subramanian, R.R.; Masters, S.C. 14-3-3 Proteins: Structure, function, and regulation. *Annu. Rev. Pharmacol. Toxicol.* **2000**, *40*, 617–647, doi:10.1146/annurev.pharmtox.40.1.617.

96. Hartman, A.M.; Hirsch, A.K.H. Molecular insight into specific 14-3-3 modulators: Inhibitors and stabilisers of protein–protein interactions of 14-3-3. *Eur. J. Med. Chem.* **2017**, *136*, 573–584.

97. Nathan, K.G.; Lal, S.K. The Multifarious Role of 14-3-3 Family of Proteins in Viral Replication. *Viruses* **2020**, *12*, doi:10.3390/V12040436.

98. Dougherty, M.K.; Morrison, D.K. Unlocking the code of 14-3-3. *J. Cell Sci.* **2004**, *117*, 1875–1884, doi:10.1242/jcs.01171.

99. Aoki, H.; Hayashi, J.; Moriyama, M.; Arakawa, Y.; Hino, O. Hepatitis C Virus Core Protein Interacts with 14-3-3 Protein and Activates the Kinase Raf-1. *J. Virol.* **2000**, *74*, 1736, doi:10.1128/JVI.74.4.1736-1741.2000.

100. Surjit, M.; Kumar, R.; Mishra, R.N.; Reddy, M.K.; Chow, V.T.K.; Lal, S.K. The Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein Is Phosphorylated and Localizes in the Cytoplasm by 14-3-3-Mediated Translocation. *J. Virol.* **2005**, *79*, 11476–11486, doi:10.1128/jvi.79.17.11476-11486.2005.

101. Tugaeva, K. V.; Hawkins, D.E.D.P.; Smith, J.L.R.; Bayfield, O.W.; Ker, D.S.; Sysoev, A.A.; Klychnikov, O.I.; Antson, A.A.; Sluchanko, N.N. The Mechanism of SARS-CoV-2 Nucleocapsid Protein Recognition by the Human 14-3-3 Proteins :SARS-CoV-2 N association with host 14-3-3 proteins. *J. Mol. Biol.* **2021**, *433*, 166875, doi:10.1016/j.jmb.2021.166875.

102. Johnson, C.; Crowther, S.; Stafford, M.J.; Campbell, D.G.; Toth, R.; MacKintosh, C. Bioinformatic and experimental survey of 14-3-3-binding sites. *Biochem. J.* **2010**, *427*, 69, doi:10.1042/BJ20091834.

103. Welker, L.; Paillart, J.-C.; Bernacchi, S. Importance of Viral Late Domains in Budding and Release of Enveloped RNA Viruses. *Viruses* **2021**, *13*, doi:10.3390/V13081559.

104. Votteler, J.; Sundquist, W.I. Virus Budding and the ESCRT Pathway. *Cell Host Microbe* **2013**, *14*, 232–241, doi:10.1016/J.CHOM.2013.08.012.

105. Freed, E.O. Viral Late Domains. *J. Virol.* **2002**, *76*, 4679, doi:10.1128/JVI.76.10.4679-4687.2002.

106. Honeychurch, K.M.; Yang, G.; Jordan, R.; Hruby, D.E. The Vaccinia Virus F13L YPPL Motif Is Required for Efficient Release of Extracellular Enveloped Virus. *J. Virol.* **2007**, *81*, 7310, doi:10.1128/JVI.00034-07.

107. Barouch-Bentov, R.; Neveu, G.; Xiao, F.; Beer, M.; Bekerman, E.; Schor, S.; Campbell, J.; Boonyaratanakornkit, J.; Lindenbach, B.; Lu, A.; et al. Hepatitis C Virus Proteins Interact with the Endosomal Sorting Complex Required for Transport (ESCRT) Machinery via Ubiquitination To Facilitate Viral Envelopment. *MBio* **2016**, *7*, doi:10.1128/MBIO.01456-16.

108. Harty, R.N.; Brown, M.E.; McGettigan, J.P.; Wang, G.; Jayakar, H.R.; Huibregtse, J.M.; Whitt, M.A.; Schnell, M.J. Rhabdoviruses and the Cellular Ubiquitin-Proteasome System: a Budding Interaction. *J. Virol.* **2001**, *75*, 10623, doi:10.1128/JVI.75.22.10623-10629.2001.

109. Shimode, S.; Nakaoka, R.; Hoshino, S.; Abe, M.; Shogen, H.; Yasuda, J.; Miyazawa, T. Identification of cellular factors required for the budding of koala retrovirus. *Microbiol. Immunol.* **2013**, *57*, 543–546, doi:10.1111/1348-0421.12066.

110. Wolff, S.; Ebihara, H.; Groseth, A. Arenavirus Budding: A Common Pathway with Mechanistic Differences. *Viruses* **2013**, *5*, 528, doi:10.3390/V5020528.

111. Dolnik, O.; Kolesnikova, L.; Stevermann, L.; Becker, S. Tsg101 Is Recruited by a Late Domain of the Nucleocapsid Protein To Support Budding of Marburg Virus-Like Particles. *J. Virol.* **2010**, *84*, 7847–7856, doi:10.1128/jvi.00476-10.

112. Harty, R.N.; Brown, M.E.; Wang, G.; Huibregtse, J.; Hayes, F.P. A PPxY motif within the VP40 protein of Ebola virus interacts physically and functionally with a ubiquitin ligase: Implications for filovirus budding. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 13871–13876, doi:10.1073/pnas.250277297.

113. Martin-Serrano, J.; Zang, T.; Bieniasz, P.D. HIV-1 and Ebola virus encode small peptide motifs that recruit Tsg101 to sites of particle assembly to facilitate egress. *Nat. Med.* **2001**, *7*, 1313–1319, doi:10.1038/nm1201-1313.

114. Iglesias-Bexiga, M.; Palencia, A.; Corbi-Verge, C.; Martin-Malpartida, P.; Blanco, F.J.; Macias, M.J.; Cobos, E.S.; Luque, I. Binding site plasticity in viral PPxY Late domain recognition by the third WW domain of human NEDD4. *Sci. Reports* **2019**, *9*, 1–17, doi:10.1038/s41598-019-50701-3.

115. VerPlank, L.; Bouamr, F.; LaGrassa, T.J.; Agresta, B.; Kikonyogo, A.; Leis, J.; Carter, C.A. Tsg101, a homologue of ubiquitin-conjugating (E2) enzymes, binds the L domain in HIV type 1 Pr55Gag. *Proc. Natl. Acad. Sci.* **2001**, *98*, 7724–7729, doi:10.1073/PNAS.131059198.

116. Rose, K.M. When in need of an ESCRT: The nature of virus assembly sites suggests mechanistic parallels between nuclear virus egress and retroviral budding. *Viruses* **2021**, *13*, doi:10.3390/v13061138.

117. Calistri, A.; Reale, A.; Palù, G.; Parolin, C. Why Cells and Viruses Cannot Survive without an ESCRT. *Cells* **2021**, *10*, 483, doi:10.3390/CELLS10030483.

118. Peng, Z.; Yan, J.; Fan, X.; Mizianty, M.J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V.N.; Kurgan, L. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **2014**, *72*, 137–151, doi:10.1007/s00018-014-1661-9.

119. Kastano, K.; Erdős, G.; Mier, P.; Alanis-Lobato, G.; Promponas, V.J.; Dosztányi, Z.; Andrade-Navarro, M.A. Evolutionary Study of Disorder in Protein Sequences. *Biomolecules* **2020**, *10*, 1–17, doi:10.3390/BIOM10101413.

120. Koonin, E. V.; Wolf, Y.I. Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.* **2010**, *11*, 487, doi:10.1038/NRG2810.

121. Tompa, P.; Szász, C.; Buday, L. Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.* **2005**, *30*, 484–489, doi:10.1016/j.tibs.2005.07.008.

122. Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2014**, *16*, 18–29, doi:10.1038/nrm3920.

123. Zhang, J.; Cruz-cosme, R.; Zhuang, M.W.; Liu, D.; Liu, Y.; Teng, S.; Wang, P.H.; Tang, Q. A systemic and molecular study of subcellular localization of SARS-CoV-2 proteins. *Signal Transduct. Target. Ther.* **2020**, *5*, 1–3.

124. Gordon, D.E.; Hiatt, J.; Bouhaddou, M.; Rezelj, V. V.; Ulferts, S.; Braberg, H.; Jureka, A.S.; Obernier, K.; Guo, J.Z.; Batra, J.; et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science (80-.).* **2020**, *370*, eabe9403, doi:10.1126/science.abe9403.

125. Boson, B.; Legros, V.; Zhou, B.; Siret, E.; Mathieu, C.; Cosset, F.-L.; Lavillette, D.; Denolly, S. The SARS-CoV-2 Envelope and Membrane proteins modulate maturation and retention of the Spike protein, allowing assembly of virus-like particles. *J. Biol. Chem.* **2021**, *296*, 100111., doi:10.1074/jbc.RA120.016175.

126. Lee, J.-G.; Huang, W.; Lee, H.; van de Leemput, J.; Kane, M.A.; Han, Z. Characterization of SARS-CoV-2 proteins reveals Orf6 pathogenicity, subcellular localization, host interactions and attenuation by Selinexor. *Cell Biosci.* **2021**, *11*, 58, doi:10.1186/s13578-021-00568-7.

127. Duart, G.; García-Murria, M.J.; Grau, B.; Acosta-Cáceres, J.M.; Martínez-Gil, L.; Mingarro, I. SARS-CoV-2 envelope protein topology in eukaryotic membranes: SARS-CoV-2 E protein topology. *Open Biol.* **2020**, *10*, doi:10.1098/rsob.200209.

128. Mészáros, B.; Sámano-Sánchez, H.; Alvarado-Valverde, J.; Čalyševa, J.; Martínez-Pérez, E.; Alves, R.; Shields, D.C.; Kumar, M.; Rippmann, F.; Chemes, L.B.; et al. Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. *Sci. Signal.* **2021**, *14*, 334, doi:10.1126/SCISIGNAL.ABD0334.

129. Kliche, J.; Kuss, H.; Ali, M.; Ivarsson, Y. Cytoplasmic short linear motifs in ACE2 and integrin β3 link SARS-CoV-2 host cell receptors to mediators of endocytosis and autophagy. *Sci. Signal.* **2021**, *14*, 1117, doi:10.1126/SCISIGNAL.ABF1117.

130. Thul, P.J.; Åkesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Blal, H.A.; Alm, T.; Asplund, A.; Björk, L.; Breckels, L.M.; et al. A subcellular map of the human proteome. *Science.* **2017**, *356*, doi:10.1126/SCIENCE.AAL3321.

131. Uhlén, M.; Fagerberg, L.; Hallström, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Tissue-based map of the human proteome. *Science.*. **2015**, *347*, doi:10.1126/SCIENCE.1260419.

132. Peacock, T.P.; Goldhill, D.H.; Zhou, J.; Baillon, L.; Frise, R.; Swann, O.C.; Kugathasan, R.; Penn, R.; Brown, J.C.; Sanchez-David, R.Y.; et al. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nat. Microbiol.* **2021**, *6*, 899–909, doi:10.1038/s41564-021-00908-w.

CHAPTER III: COMPARATIVE ANALYSIS OF STRUCTURAL FEATURES IN SLIMS FROM EUKARYOTES, BACTERIA, AND VIRUSES WITH IMPORTANCE FOR HOST-PATHOGEN INTERACTIONS

ABSTRACT

Protein-protein interactions drive functions in eukaryotes that can be described by short linear motifs (SLiMs). Conservation of SLiMs help illuminate functional SLiMs in eukaryotic protein families. However, the simplicity of eukaryotic SLiMs makes them appear by chance due to mutational processes not only in eukaryotes but also in pathogenic bacteria and viruses. Further, functional eukaryotic SLiMs are often found in disordered regions. Although proteomes from pathogenic bacteria and viruses have less disorder than eukaryotic proteomes, their proteins can successfully mimic eukaryotic SLiMs and disrupt host cellular function. Identifying important SLiMs in pathogens is difficult but essential for understanding potential host-pathogen interactions. We performed a comparative analysis of structural features for experimentally verified SLiMs from the Eukaryotic Linear Motif (ELM) database across viruses, bacteria, and eukaryotes. Our results revealed that many viral SLiMs and specific motifs found across viruses and eukaryotes, such as some glycosylation motifs, have less disorder. Analyzing the disorder and coil properties of equivalent SLiMs from pathogens and eukaryotes revealed that some motifs are more structured in pathogens than their eukaryotic counterparts and vice versa. These results support a varying mechanism of interaction between pathogens and their eukaryotic hosts for some of the same motifs.

Introduction

Protein-protein interactions (PPIs) are pivotal for modulating intracellular processes [1,2]. PPI networks are often regulated through transient interactions, mediated by unstructured protein regions that lack a well-defined structural conformation [3–5]. Alteration of PPI networks inside the cell can trigger disease [1,2]. Protein-protein interactions are often mediated by short linear motifs (SLiMs) [6–8]. SLiMs are short sequence patterns, with an average length ranging from 3 to 10 sequential residues [5,9,10]. SLiMs can be represented by regular expressions that describe the evolvability of the sequence pattern where amino acid replacements may occur at specific positions, while other positions must be strictly conserved to ensure functionality [5,9,10]. Given the simplicity (the short length and high evolvability) of SLiM sequence patterns, they may occur by chance [11]. When proteins from pathogens display a sequence pattern that matches SLiM motifs in their host, molecular mimicry can result. Through molecular mimicry of SLiMs, pathogen proteins can disrupt native host interactions, often to the benefit of the pathogen. SLiMs are pathogens' vehicle to hijack and rewire the host interactome [5,8].

In 2007, Fuxreiter and coworkers showed that verified SLiMs from the Eukaryotic Linear Motifs (ELM) database were predicted to be mostly intrinsically disordered [10]. The SLiMs were present in more disordered regions with respect to their global surrounding sequence based on disorder prediction. However, the SLiMs themselves, although still disordered, were found to be slightly less disordered than their local adjacent sequence [10]. Intrinsic disorder in proteins refers to a multi-conformational structure with high plasticity and an ability to fold and unfold [12]. The amount of

intrinsic disorder can range from small protein regions (intrinsically disordered regions: IDRs) to fully disordered protein (intrinsically disordered proteins: IDPs) [13]. The inherent conformational plasticity of intrinsic disorder allows for conformational changes, which can induce a local structure transition that is essential for a successful protein-protein interaction [12,13].

A hurdle in SLiM discovery is the high false-positive rate associated with computational identification approaches. Searching for a SLiM motif using only regular expressions can lead to the discovery of many instances by chance [9,14]. A common approach of reducing false-positive SLiMs is to exclude matches that are not intrinsically disordered [9,15]. However, other studies showed that disorder is not always necessary for the functionality of the motifs [16]. False-positives may also be removed by considering the conservation of the SLiM across homologous proteins [17,18]. While this may work for eukaryotic SLiMs that are more conserved [5], it can prevent the identification of functional SLiMs occurring by chance in regions with high evolutionary rates, such as SLiMs in IDRs [11]. Further, considering only conserved motifs as functional can fail to remove false-positive results occurring in a conserved globular region [19]. Via and coworkers proposed that consideration of surface accessibility and susceptibility to be a loop (to not fold into a secondary structure) could improve the identification of true positive SLiMs, but it may discard buried SLiMs that can be accessible due to allosteric effects [18].

We recently analyzed the viral SLiMs from the ELM database, and many of the experimentally verified functional SLiMs were discovered to be devoid of disorder [20]. The lack of disorder in some viral SLiMs is not surprising due to the low disorder content

in some viral families [21]. Moreover, eukaryotic proteins include a higher percentage of disordered protein regions than bacteria and viruses [22,23]. Currently, the ELM database contains almost 4,000 experimentally verified instances from approximately 300 motifs in proteins from eukaryotes, bacteria, and viruses. SLiMs in the ELM database are divided into six functional categories: cleavage motifs (CLV), degradation motifs (DEG), docking motifs (DOC), ligand-binding motifs (LIG), post-translational modification motifs (MOD), and targeting motifs (TRG) [24]. Due to the growth of the ELM database, reanalysis of SLiMs is essential to highlight differences and similarities between eukaryotes, bacteria, and viruses. Further, a comparison of viral and bacterial SLiMs with their eukaryotic counterparts is warranted to disentangle whether potential differences stem from the taxonomic group or ELM functionality. To this end, we present a comparative analysis of SLiMs from eukaryotes, bacteria, and viruses based on sequence-based predictions of structural characteristics to identify similarities and differences between known SLiMs.

Results and Discussion

1.1. The Majority of Instances in the ELM Database Bind Ligands and are from Human

From the 3934 instances downloaded from the ELM database, 3716 were annotated as true positive SLiMs. Only true positive SLiMs were analyzed, and from hereinafter, SLiMs refer to true positive SLiMs. SLiMs from viruses are often listed as their viral polyprotein product and not as the processed functional protein. Polyproteins can impact the predictions of structural features and thus, we used only viral proteins for which the

functional protein product could be determined. Three viral instances were excluded from further analysis since their functional protein could not be resolved.

In the final dataset of SLiMs used here, the majority of the instances are from eukaryotes (3320 instances), followed by viruses (278 instances) and bacteria (115 instances) (Figure 1A). Most eukaryotic SLiMs are from vertebrates, specifically Homo sapiens (2056). The major type of SLiMs represented in all groups is the LIG binding motifs with 1839 SLiMs and the composition of other types varies by taxonomic group (Figure 1B). Although the low number of instances from other taxonomic groups reduces the power of any comparative analysis, to better understand the landscape of SLiMs, we characterize the structural properties associated with the function type of SLiMs in general and specifically compare SLiMs across taxonomic groups when possible.

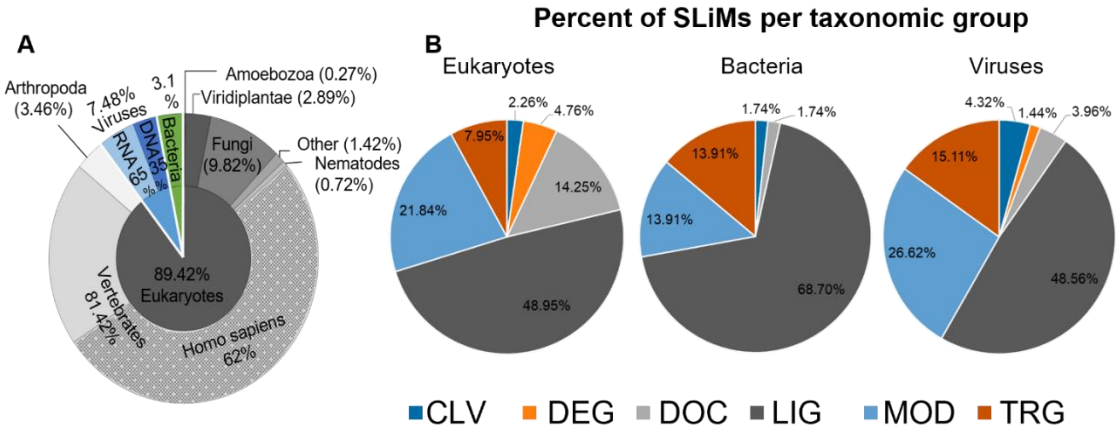

Figure 1. The SLiM dataset composition by taxonomy and functionality. The percentage of SLiMs per taxonomic group and taxonomic subgroup; eukaryotes and its subgroups (grey), viruses and its subgroups (blue), and bacteria (green) based on all SLiMs (A). The percentage of SLiMs is colored by functional type in each taxonomic group (B). For further information, see Table S1.

1.2. Accessibility and Lack of Secondary Structure Influence SLiM Functionality More Than Disorder

To analyze the structural properties of all SLiMs, we predicted intrinsic disorder, surface accessibility, and secondary structure for all residues in proteins with SLiMs in our dataset. The predictions were used to classify each residue within a SLiM as i) either disordered or ordered based on a cutoff, ii) either accessible or buried based on a cutoff, and iii) either in a secondary structure or not in a secondary structure (hereinafter referred to as coil) based on probability from the prediction. The classifications were used to calculate the percentage of disorder, accessibility, and coil, respectively, for each SLiM. Using the default IUPRED2A cutoff (0.5) and the long disorder option to infer disorder, eukaryotes (46%) and bacteria (53%) have a greater proportion of fully disordered instances than viruses (39%) (Figure 2A). Viruses (44%) have a higher percentage of fully ordered instances than eukaryotes (31%) and bacteria (25%). The analysis was repeated with the 0.5 cutoff using the IUPRED2A short disorder prediction option, and the overall trends are the same (Figure 2C). While the IUPRED2A default cutoff for disorder is 0.5, a lower cutoff (0.4) is often used to assign disorder [10]. Using the lower cutoff (0.4), more disordered and less ordered instances were observed with both the long and short IUPRED2A prediction (Figure 2B, 2D). Few instances have a mixture of ordered and disordered residues for all three groups at both cutoffs (Figure 2A-D). The disorder per SLiM changes for some SLiMs from the long to short IUPRED2A prediction, but the overall correlation is high (Figure S1). In eukaryotes and viruses, the Spearman correlation is high (r_s=0.77 and 0.76, and the *p*-value is 0 and 1.30×10^{-54}, respectively), while in bacteria, a moderate correlation was observed (r_s = 0.52, *p*-value =

1.31×10^{-9}). For accessibility, we found that eukaryotes (77%) and bacteria (76%) have an increased share of fully accessible instances compared to viruses (71.5%). The remaining percentages vary in percent accessibility per instance in all taxonomic groups (Figure 2E). For coil, eukaryotes and bacteria are similar, with approximately 80% of their instances predicted to be coil (not alpha helix or beta-strand), compared to 68% for viruses. Less than 7% of viral SLiMs and almost 5% of both eukaryotic and bacterial SLiMs were found to have secondary structures (Figure 2D).

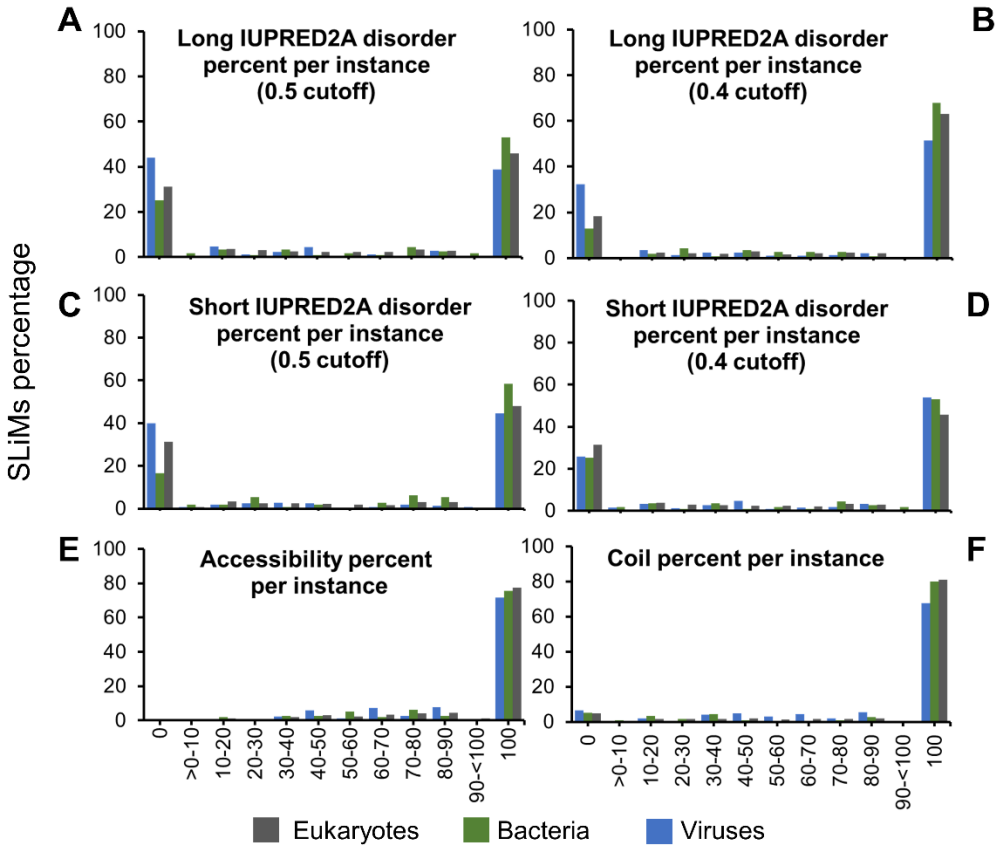

Figure 2. Predicted properties per instance across taxonomic groups. The predicted percentage per instance; IUPRED2A long disorder based on 0.5 cutoff (A) and0.4 cutoff (B), IUPRED2A short disorder based on 0.5 cutoff (C) and 0.4 cutoff (D), NetSurfP 2.0 accessibility based on 0.25 cutoff (E), and NetSurfP 2.0 prediction of coil based on three state analysis (F). For further information, see Table S1.

These results reveal that while disorder content varies greatly, accessibility and coil content are prevalent properties across the SLiM distribution. Altogether, these findings suggest a large impact on the functionality of SLiMs for the latter two and an interplay between order and disorder with accessibility and coil. By being fully accessible, SLiMs can interact with other proteins. For partially accessible instances, critical amino acids required for the interaction may be the only exposed residues. Alternatively, the SLiM may be fully or partially concealed until the proper cellular conditions contribute to changing its conformation to become accessible for the interaction to occur. Thus, partially accessible SLiMs could play a pivotal role in regulating the functional cascade triggered by a SLiM. Coil and disorder predictions indicate dynamic, flexible structures for which the conformational ensemble population can vary due to the cellular environment, affecting functional conformations to various degrees. It is plausible that conformational flexibility varies by functionality, such as ELM type. Further, these binary classifications simplify the predictions as a percentage per instance and may not reveal important information about the SLiMs attributes. Exploring the mean IUPRED2A disorder score and mean coil confidence score of SLiMs in each taxonomic category and by ELM type can provide more insights into their structural and functional properties.

1.3. SLiMs from Viruses are Less Disordered

To explore the mean IUPRED2A disorder score (MIDS) of SLiMs by ELM type, the IUPRED2A disorder prediction scores for all residues within a SLiM were averaged for both long and short disorder predictions, respectively. There is good agreement between long and short disorder prediction overall, but important shifts towards higher disorder

from long to short are observed for certain ELM types (Figure S2-S4). A strong positive correlation between long and short disorder predictions that is statistically significant was observed in most of the comparisons by ELM type in each taxonomic group. However, some showed either weak positive correlation, such as DEG in eukaryotes (r_s = 0.37, *p*-value = 1.34×10^{-6}), or moderate correlation such as TRG in eukaryotes (r_s = 0.66, *p*-value = 1.26×10^{-35}) and LIG and TRG in viruses ((r_s = 0.60 and 0.66, *p*-value = 8.8×10^{-15} and 1.74×10^{-6}, respectively), or no correlation such as MOD in bacteria (r_s = 0.03, *p*-value = 0.89). Some motif instances shift from ordered to disordered from the long to the short IUPRED2A prediction, suggesting that not all instances are found in long disordered regions but in short disordered loops. Hence, using only the default long disorder prediction for short viral and bacterial proteins that are known to lack long disordered domains may impact the disorder content of SLiMs and lead to the exclusion of functional motifs in non-eukaryotic pathogens.

Hypothesis testing was performed to compare the MIDS distribution for all instances between different ELM types and taxonomic groups. The MIDS values vary greatly between ELM types. For MIDS based on long disorder, MOD and TRG have lower MIDS values than LIG and DOC (Figure 3A). For MIDS based on short disorder, MOD is lower than LIG and TRG, and CLV is lower than DEG (Figure 3C).

Additional MIDS analysis was performed within taxonomic groups by ELM type to investigate differences across taxonomic groups. While most comparisons in both long and short IUPRED2A MIDS are not significant due to the wide dispersion of data within each ELM type, some are significant and may provide insights into the discrepancy of MIDS between ELM types (Figure 3B-3D). In viruses, both long and short IUPRED2A

MIDS analysis showed that LIG and TRG have higher MIDS values than MOD (adjusted *p*-value = 0 and 7×10^{-6}, respectively). In bacteria, for the long IUPRED2A MIDS, only LIG is higher than MOD (adjusted *p*-value = 3.86×10^{-2}). However, there were more observed differences between ELM types in the eukaryotic instances, especially for long IUPRED2A. For DOC, MIDS values are higher than MOD and TRG (adjusted *p*-value = 2.20×10^{-3} and 0, respectively) for long IUPRED2A and LIG is higher than MOD and TRG motifs ((adjusted *p*-value = 1.85×10^{-2} and 4.26×10^{-4} respectively) (Figure 3B). For short IUPRED2A MIDS, eukaryotes LIG and DEG were higher than MOD (adjusted *p*-value = 4.41×10^{-2}, and 3.19×10^{-3}). The increased MIDS values of the long disorder prediction for DOC and LIG, the main motif types that involve interaction with other proteins inside the cell, support their dynamic role in regulating cellular pathways and machinery and suggests their presence in long disordered regions rather than short disordered loops. Comparing MIDS for the same ELM type across taxonomic groups revealed that the only difference between taxonomic groups was for MOD motifs in viruses which had lower MIDS than eukaryotes (adjusted *p*-value = 0) for both long and short disorder, and bacteria (adjusted *p*-value = 8.52×10^{-4}) for short disorder.

To further explore the distribution of MIDS by taxonomic group, the percentage of SLiMs per ELM type across different MIDS ranges were plotted (Figure 3E-J). Based on long disorder analysis of the MIDS values and a 0.4 cutoff, approximately 73% and 77% of the instances in eukaryotes and bacteria, respectively, are disordered, while approximately 59% of the viral instances are disordered. Proportionally, the amount of each ELM type appears similar across MIDS bins for all taxonomic groups, except MOD in bacteria and viruses, which have lower MIDS values (Figure 3E-F). For the short

IUPRED2A disorder analysis of MIDS, the amount of disordered eukaryotic SLiMs is similar to MIDS from long disorder, while viruses and bacteria show an increased amount of disordered SLiM for MIDS from short disorder (64% and almost 83%, respectively). Most MOD instances from viruses and bacteria have lower MIDS values than eukaryotes for long IUPRED2A prediction (Figure 3E-F). For short IUPRED2A prediction, MOD in bacteria is more disordered, and viruses show a subtle shift towards disorder for some instances (Figure 3I-J). Notably, the number of instances in the highest MIDS category based on long disorder is reduced for bacteria and viruses for short disorder (Figure 3I-J). An analysis of MODs from the same motifs from different taxonomic groups is required to generalize or discard this trend.

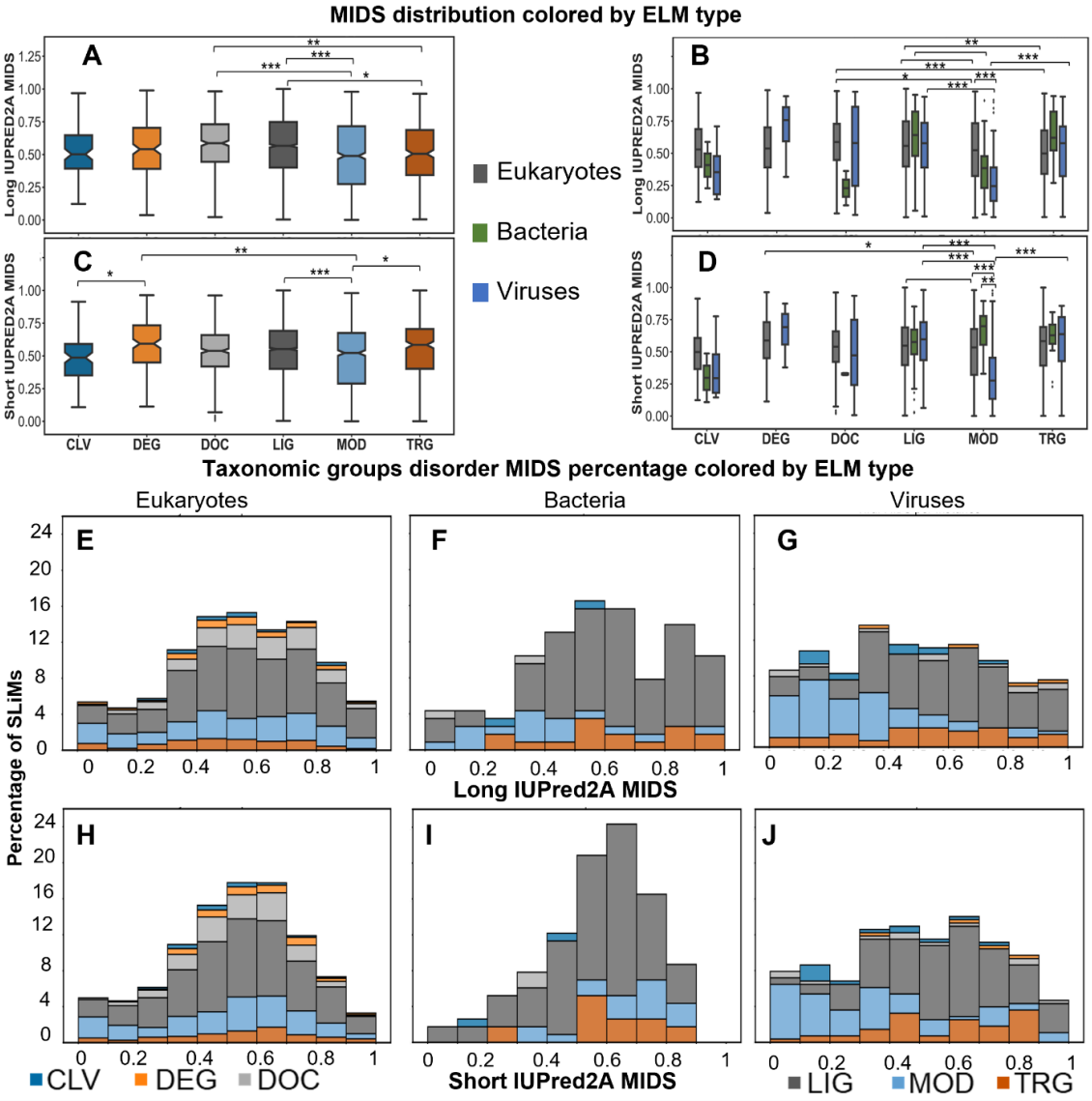

Figure 3. Distribution of MIDS values. Boxplots for the distribution of long IUPRED2A MIDS of all SLiMs per motif type colored as shown by legend (A). Boxplots for long IUPRED2A MIDS distribution of all SLiMs in each taxonomic group (bacteria (green), viruses (blue), eukaryotes (grey)) classified based on their ELM type (B). Boxplots for the distribution of long IUPRED2A MIDS of all SLiMs per motif type colored as shown by legend (C). Boxplots for long IUPRED2A MIDS distribution of all SLiMs in each taxonomic group, colored as in (B), classified based on their ELM type (D). Hypothesis testing with Mann-Whitney test with simple Bonferroni correction was performed and significant adjusted *p*-values in (A) and (B) are shown as brackets between groups (No asterisk for adjusted *p*-values between 0.05 and <0.01, * for adjusted *p*-value ≤ 0.01, ** for $\leq 1 \times 10^{-3}$, and *** for $\leq 1 \times 10^{-4}$). The sample size per each tested group and adjusted *p*-values can be found in Table S1. The percentage of SLiMs by long IUPED2A MIDS range in different taxonomic groups colored by ELM type (E-G). The

percentage of SLiMs by short IUPED2A MIDS range in different taxonomic groups colored by ELM type (H-J), colored as in (A). For more information, see Tables S1 and S2.

1.4. Most SLiMs Lack Secondary Structure

To explore the mean coil confidence score (MCCS) of SLiMs by ELM type, the NetSurfP coil confidence scores for all residues within a SLiM were averaged. Hypothesis testing was performed to compare the MCCS distribution for all instances between different ELM types and taxonomic groups. All data have a negatively skewed distribution with the highest percentages of SLiMs in the upper bin range of 0.9 to 1 MCCS values. Analyzing the SLiMs MCCS by ELM type revealed that DOC has higher MCCS values than CLV and MOD (adjusted *p*-value = 1.25×10^{-2} and 7.1×10^{-3}, respectively). DEG also shows an increase in MCCS values compared to CLV, and MOD types (adjusted *p*-value = 3.19×10^{-3}, and 6.24×10^{-3}, respectively) (Figure 3A). Intrinsic disorder properties of proteins have previously been linked to proteins being unstructured or having enough plasticity to undergo structural transitions [25,26]. While most SLiMs are predicted to have high coil confidence, instances from some ELM types show great variation in MCCS values. This may indicate a presence of structural transitions and spatiotemporal control of the structure to perform the function of the SLiMs, but it may also indicate that some SLiMs are not conformationally flexible but lack disorder and have secondary structure.

Instances analysis of MCCS across taxonomic groups highlighted great variability in CLV and LIG for all groups. In MOD only viruses showed great variability. In viruses, MCCS values for MOD motifs were lower than LIG and TRG (adjusted *p*-value = 0 and 2.27×10^{-3}, respectively). Comparing scores for MOD between taxonomic groups found

viruses lower than eukaryotes and bacteria (adjusted *p*-value = 0 and 5.04×10^{-3}, respectively) (Figure 4B).

To further explore the distribution of MCCS by taxonomic group, the percentage of SLiMs per ELM type across different MCCS ranges were plotted (Figure 4C-E). The analysis of SLiMs percent distribution of MCCS in all taxonomic groups revealed that most SLiMs have high coil confidence (Figure 4C-E). Approximately 80% in eukaryotes, 60% in bacteria, and 50% in viruses have MCCS > 0.9. The LIG motifs are the predominant motifs with MCCS > 0.9. MOD sites have higher distribution over all MCCS ranges of viral instances than other taxonomic groups. Altogether, the lower values and the great variability in MCCS of viral MOD sites support the MIDS results, suggesting that some modification sites, especially in viruses, are ordered (not disordered or coil).

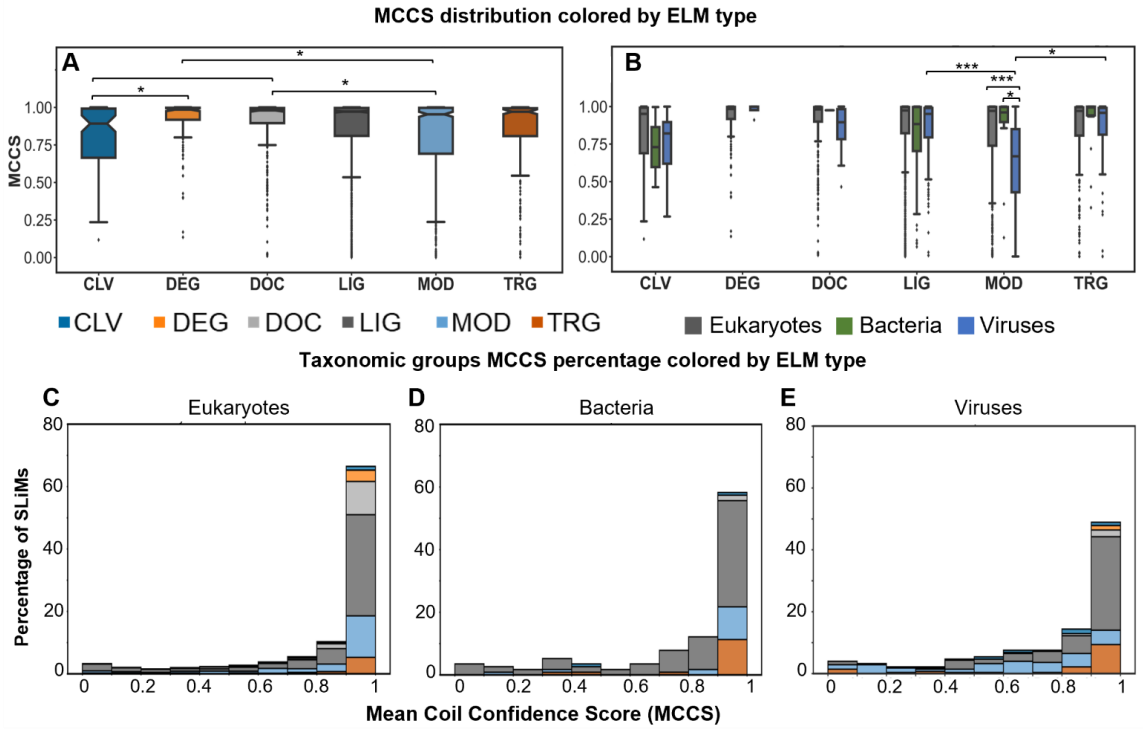

Figure 4. Distribution of MCCS values. Boxplots for the distribution of MCCS of all SLiMs per motif type colored as shown by legend (A). Boxplots for MCCS distribution of all SLiMs in each taxonomic group (bacteria in green, viruses in blue, and eukaryotes in grey) classified based on their ELM type (B). Hypothesis testing with Mann-Whitney test with simple Bonferroni correction was performed and significant adjusted *p*-values in (A) and (B) are shown as brackets between groups (No asterisk for adjusted *p*-values between 0.05 to <0.01, * for adjusted *p*-value ≤ 0.01, and *** for $\leq 1 \times 10^{-4}$). The sample size per each tested group and adjusted *p*-values can be found in Table S1. The percentage of SLiMs by MCCS range in different taxonomic groups colored by ELM type (C-E) colored as in (A). For more information, see Tables S1 and S2.

1.5. Disordered or Flexible?

The above-mentioned results led us to pose three questions. First, we asked whether SLiMs possess intrinsic disorder and coil properties that differ from of the overall protein context. Second, we asked if the same SLiMs from different taxonomic groups are different from each other? Third, for the shared motif instances, is there any variation in their structure-based sequence properties that might affect the functionality between different groups?

2.5.1. SLiMs are Found in Flexible Regions

To answer the first question of whether SLiMs differ in intrinsic disorder content and in secondary structure compared to the overall protein context, we extracted the long and short disorder and coil confidence scores for the flanking regions of each instance. We examined the 100 residues before and after the SLiM instance in each taxonomic group. The mean of all positions and the 95% confidence interval were computed and plotted with the center (zero) representing the mean long or short MIDS or MCCS values per instance for all instances, based on long mean MIDS (mMIDS), short mMIDS, and mean MCCS (mMCCS), respectively (Figure 5A, C).

For disorder, instances from the three taxonomic groups show the same overall trend with increasing MIDS values towards mMIDS in both long and short disorder IUPRED2A predictions, but with less disordered flanking regions in viruses than in eukaryotes and bacteria (Figure 5A). Bacterial instances are more disordered than eukaryotic instances, however, due to the wide MIDS confidence interval range for the bacterial instances, this might not hold true if more data are explored. Viral instances have the lowest mMIDS values between tested groups using both long and short disorder predictions, supporting previous findings that viruses can be more ordered than other taxonomic groups [22,23]. In addition, using the long IUPRED2A disorder the bacterial and eukaryotic SLiMs are located in a less disordered region than their immediate surrounding region (Figure 5A), in agreement with previous work [10]. The effect is more prominent in eukaryotes, where a crater-like dip in disorder surrounds the mMIDS and may change for bacteria if more data was available. When SLiMs are located in a less disordered protein region than the surrounding region, the highly disordered surrounding regions can regulate or enhance the binding of the SLiM in protein complexes in accordance with protein fuzziness [27,28]. We observe no such pattern of a dip in mMIDS for SLiMs compared to the flanking region in viruses. Moreover, no dip was observed when using the short IUPRED2A prediction, where mMIDS for all taxonomic groups are in an overall more disordered region than the flanking region (Figure 5A-5C). It should be noted that due to limited data availability for both virus and bacterial instances in the ELM database, the confidence intervals in viral and bacterial results have higher uncertainty than in eukaryotes.

The density curve for long disorder MIDS per SLiM per taxonomic group demonstrates a higher density above the cutoff 0.4 for eukaryotes and bacteria while viruses reveal almost equal density for the entire range. For both eukaryotes and bacteria, the density becomes more centered around the median for the short disorder MIDS per SLiM, with the highest density of SLiMs found at approximately 0.6 IUPRED2A short disorder value. For viruses, two subtle peaks of high density of instances appeared at approximately 0.4 and 0.6 for the short disorder MIDS per SLiM (Figure 5B-5D).

For all taxonomic groups, MCCS of flanking regions increases towards mMCCS of SLiMs (Figure 5E). SLiMs from eukaryotes and bacteria have relatively similar mMCCS values, while SLiMs from viruses have lower mMCCS values. The flanking regions in viruses have lower MCCS than eukaryotes and bacteria (Figure 5C). The density curve for MCCS per SLiMs per taxonomic group shows that viruses closely resemble bacteria above 0.8. The corresponding density plot for eukaryotes has a high density near the maximum MCCS (Figure 5D).

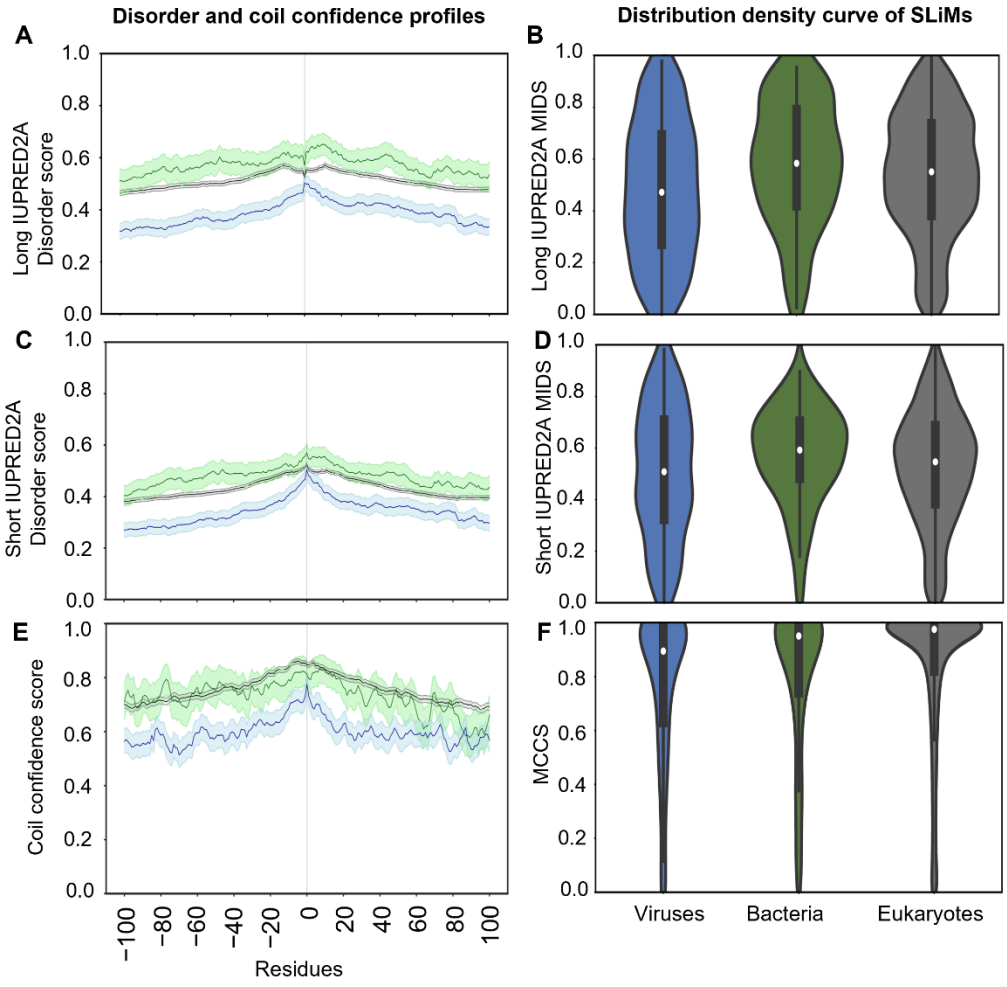

Figure 5. Disorder and coil confidence profiles of proteins containing SLiMs and the density curve of MIDS and MCCS of SLiMs per taxonomic group. The flanking regions of 100 residues around SLiMs using long IUPRED2A disorder score per taxonomic group and the 95% confidence interval of the mean (A). SLiMs long IUPRED2A MIDS density distribution plot of the SLiMs per taxonomic group (B). The flanking regions of 100 residues around SLiMs using short IUPRED2A disorder score per taxonomic group and the 95% confidence interval of the mean (C). SLiMs short IUPRED2A MIDS density distribution plot of the SLiMs per taxonomic group (D). The flanking regions of 100 residues around SLiMs coil confidence score per taxonomic group and the 95% confidence interval of the mean (E). SLiMs MCCS density distribution plot of the SLiMs per taxonomic group (F). For further information, see Table S3.

2.5.2. A Comparison of Viral and Bacterial Motifs with their Corresponding Eukaryotic Motifs

To answer the second question, if the same SLiMs from different taxonomic groups are different from each other, a correlation analysis of the disorder scores or coil confidence for SLiMs from one taxonomic group versus the equivalent SLiMs in another taxonomic group was performed. Motifs shared between taxonomic groups were extracted to compare corresponding sequence-based structural properties. Viruses and bacteria share only 16 motifs, and no further analysis was performed due to the low number of instances. Viruses and eukaryotes share 56 motifs. Bacteria and eukaryotes share 33 motifs (For detailed information see Table S4). Due to differences in the number of instances between groups for each motif, the mMIDS (long/short) and mMCCS for all instances of a motif were calculated and used to infer the correlation between shared motifs with Spearman correlation analysis. The correlation analysis for the shared motifs revealed a moderate or strong positive correlation with a significant *p*-value for all tested pairs (Figure 6).

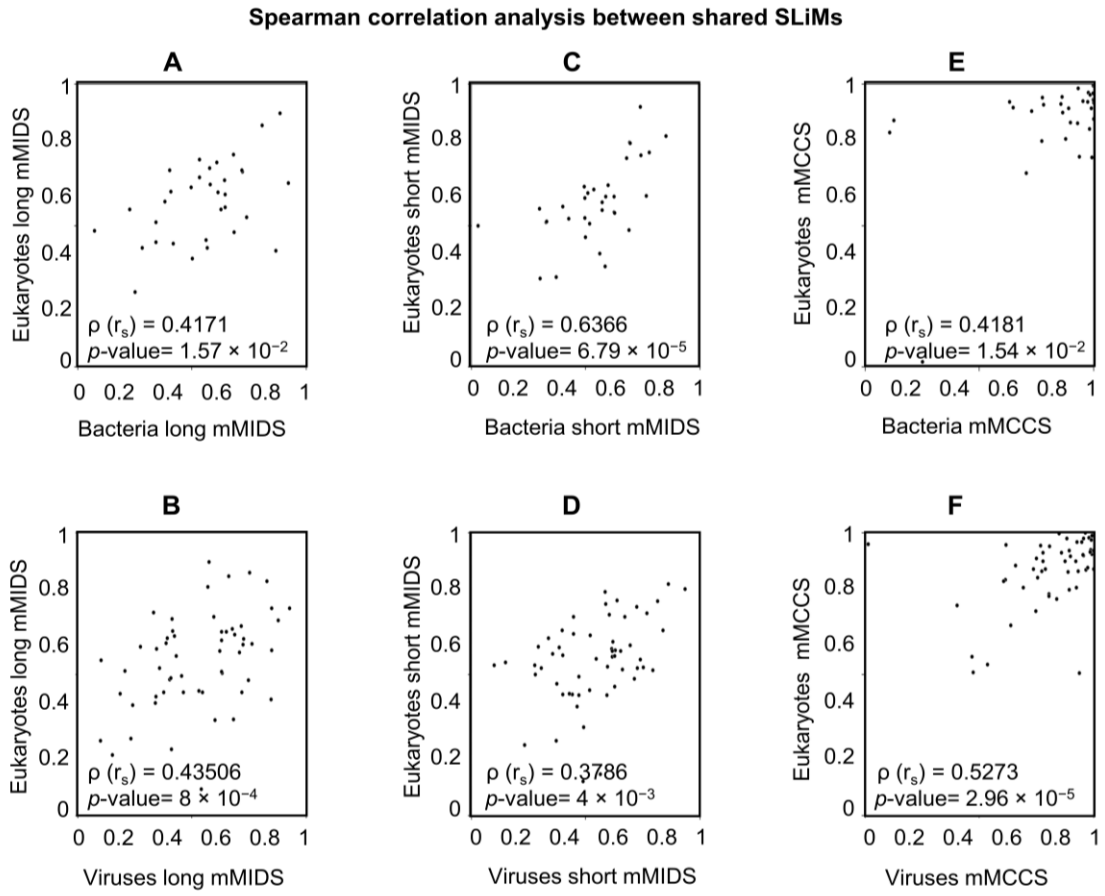

Figure 6. Scatter plot for the MIDS and MCCS means of the shared SLiMs between different groups. Long disorder MIDS means scatter plot and Spearman correlation with the *p*-value for shared SLiMs between eukaryotes vs. bacteria (A) and eukaryotes vs. viruses (B). Short disorder MIDS means scatter plot and Spearman correlation with the *p*-value for shared SLiMs between eukaryotes vs. bacteria (C) and eukaryotes vs. viruses (D). MCCS means scatter plot and Spearman correlation with the *p*-value for shared SLiMs between eukaryotes vs. bacteria (E) and eukaryotes vs. viruses (F). For detailed information about the number of instances, long/short mMIDS and mMCCS of all instances per motif, long/short MIDS and MCCS per instance, and the individual amino acid scores of disorder and coil confidence per instance, see Table S4.

The mMIDS for most shared motifs are in good agreement between the compared groups (eukaryotes and viruses or bacteria) (Table S4). For the shared motifs that were not in good agreement, the individual MIDS and MCCS of each instance and the disorder/coil confidence score per residue were inspected. Some motifs showed a

considerable variation in the MIDS and MCCS values of instances and the individual amino acid disorder and coil confidence scores. The variability in disorder scores across a motif has previously been explained by the functionality of each residue within the motif [10]. Eukaryotic motifs' wide range of MIDS and MCCS values may be influenced by numerous factors, such as the species and protein where the SLiM is found, its potential interacting protein partner, the proposed function of SLiM (to regulate the function or to activate or inhibit the function of the interacting protein permanently), the dynamics of the interaction or the interacting context in which the SLiM-protein interaction occurs (i.e., the energy of the interaction of the motif and the surrounding sequence with the interacting protein). The variability in MIDS and MCCS scores and the existence of different factors affecting SLiM interaction supports a dynamic nature of SLiMs interactions with their target protein in real-time and on evolutionary time scales as well as mutational processes.

2.5.3. To Fold or Not To Fold: A Tale of Two Motifs

To investigate our third question about the shared motif instances, is there any variation in their structure-based sequence properties that might affect the functionality between different groups? Two shared motifs between viruses and eukaryotes, MOD_N-GLC_1 and LIG_Rb_LxCxE_1, were selected for further analysis. The first motif, MOD_N-GLC-1, makes up almost 80% of all viral MOD motifs, and it is the most abundant ELM type in viruses below the 0.4 cutoff in both long and short IUPRED2A results. Overall, this motif is devoid of disorder, with long mMIDS below 0.4 for both eukaryotes and viruses, 0.27 and 0.23, respectively, and nearly similar values for the short mMIDS values as well (Table S4). The second motif, LIG_Rb_LxCxE_1, makes

up approximately 10% of all viral LIG motifs and is the most abundant LIG ELM type in viruses below the 0.4 cutoff in both disorder prediction types. The long IUPRED2A mMIDS for viruses (0.37) suggests more ordered instances, while the long IUPRED2A mMIDS for eukaryotes (0.43) suggests more disordered instances. However, the short IUPRED2A mMIDS data for viruses showed a slightly higher value of 0.42, and eukaryotes had almost equivalent value to the long IUPRED2A mMIDS value (Table S4), although this differentiation is not meaningful as all are in a similar range. Both motifs had a considerable number of instances in viruses and eukaryotes that enabled further sequence and structure investigation to discover potential differences or similarities between these two groups.

Are MOD_N-GLC_1 Instances Indeed Predominantly Ordered in Viruses or is this perhaps Due To Insufficient Data?

The MOD_N-GLC-1 motif has the regular expression pattern .(N)[^P][ST].. in the ELM database of where (dot) means any amino acid is accepted at this position, (N) means only asparagine is accepted, [^P] means any amino acid except proline is accepted and [ST] means that only serine or threonine are accepted at this position [24]. Oligosaccharyl transferase recognizes the pattern and results in N-linked glycosylation on the asparagine residue (N) at the beginning of the motif in unfolded proteins [24,29,30]. Glycosylation is a post-translational modification that usually aids in protein folding. The glycosylated protein region may acquire a specific fold, or be a part of the structured domain, or remain a coil [31]. Viral proteins are glycosylated by the glycosylation enzymes of their host [30]. Glycosylation has a wide range of effects on viruses, such as altering viral protein folding and function, inducing interactions with glycan-binding

proteins, assisting immune cell evasion, pathogenicity, cellular tropism, and blocking access to other functional regions (reviewed in [30]).

There were 156 MOD_N-GLC_1 instances in the ELM database, 59 from viruses and 97 from eukaryotes. Further analysis was performed for the distributions of long and short MIDS, and disorder score values, MCCS values, and coil confidence values per amino acid residue. The comparisons revealed no difference between eukaryotes and viruses (Figure 7). For both groups, the variation in coil confidence reaches from 0 to 1 with an accumulation at both ends. The range of coil confidence and the low disorder score of this motif may be due to the glycosylation effects. Glycosylation occurs in the endoplasmic reticulum co- or post-translationally and may induce folding of these specific sites in the protein [32–35]. Although the MOD_GLC-1_N motif instances shared between eukaryotes and viruses are often ordered, some are found to be disordered. One of the MOD_GLC-1_N motifs that is predicted to be disordered is the West Nile virus motif which also have an annotated 3D structure in the ELM database.

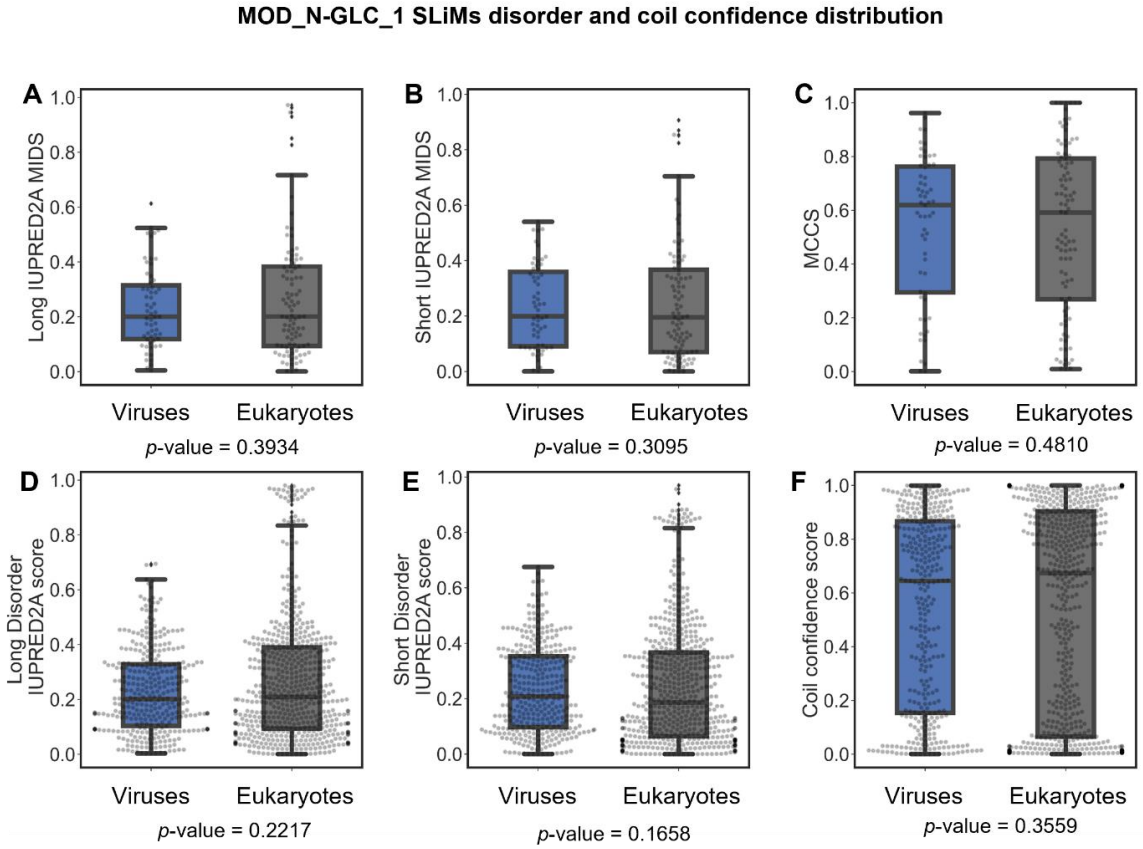

Figure 7. Disorder score and coil confidence distributions in viruses and eukaryotes for the MOD_N-GLC_1 motif. Boxplots and swarm plot distribution for SLiMs long IUPRED2A MIDS (A), short IUPRED2A MIDS (B), MCCS (C), the individual long IUPRED2A disorder scores per residue for SLiMs (D), the individual short IUPRED2A disorder scores per residue for SLiMs (E), and the individual coil confidence scores per residue for SLiMs (F).

The annotated structure from the ELM database is for the West Nile Virus (WNV) envelope protein with the MOD_N-GLC_1 motif in region 443 to 448 in PDB ID: 2HG0 [36]. Based on reduced DSSP [37] assignments of this structure, the secondary structure for the motif region is a mixture of coil and helix (CCHHHH) (Figure 8). This is different from the prediction for this instance, which is 100% coil with an MCCS value of 0.81, a MIDS score of 0.52 for long IUPRED2A prediction, and a MIDS score of 0.45 for short IUPRED2A prediction. In another structure of the same protein but without the

glycosylation (PDB ID: 3I50), this site is not resolved in the structure [38]. Unresolved, missing residues in structures from X-ray crystallography indicate disordered regions [39–41]. The MOD_N-GLC_1 motif in the envelope protein from WNV, which that is disordered when not glycosylated but adopts structure when glycosylated,flexible and accessible in the nascent initially unfolded state facilitates glycosylation by the host's Oligosaccharyl transferase. Upon glycosylation and the protein folding, the motif transitions to a more rigid (less flexible) conformation. This example illustrates how the structural states are context-dependent.

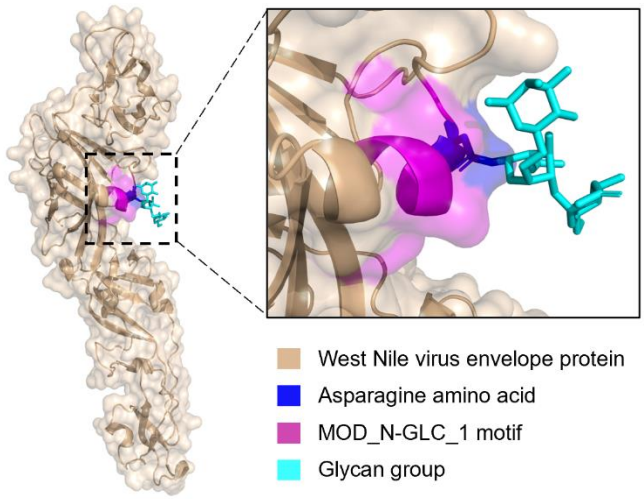

Figure 8. The glycosylated MOD_N-GLC_1 site in West Nile virus envelope protein. West Nile Virus envelope protein (beige) (PDB ID: 2HG0) rendered as a transparent surface. A closer view of the local helical structure of the MOD_N-GLC_1 motif (magenta). The glycosylated asparagine residue (blue) and glycan group (cyan) are shown as sticks.

A small phylogeny was constructed for the envelope from WNV and its homologs. Sequence-based structure properties predictions were performed and mapped to the multiple sequence alignment. The N-glycosylation site from the MOD_N-GLC_1 motif is mostly conserved, but the whole motif is missing from two close relatives of

WNV and from the Yellow Fever virus (YFV) outgroup (Figure 9). For the viruses that harbor the MOD_N-GLC_1 motif in this region, all display similar patterns of accessibility and coil, but the amount of disorder in the region varies. Although WNV envelope protein is the only annotated motif of this type in the ELM database, other viruses such as Zika, Dengue, and Japanese Encephalitis viruses were found to be glycosylated at the same alignment site and its glycosylation has been found to increase infectivity [42–47]. It is plausible that variation in flexibility of this region can impact glycosylation between viruses and consequently, their infectivity.

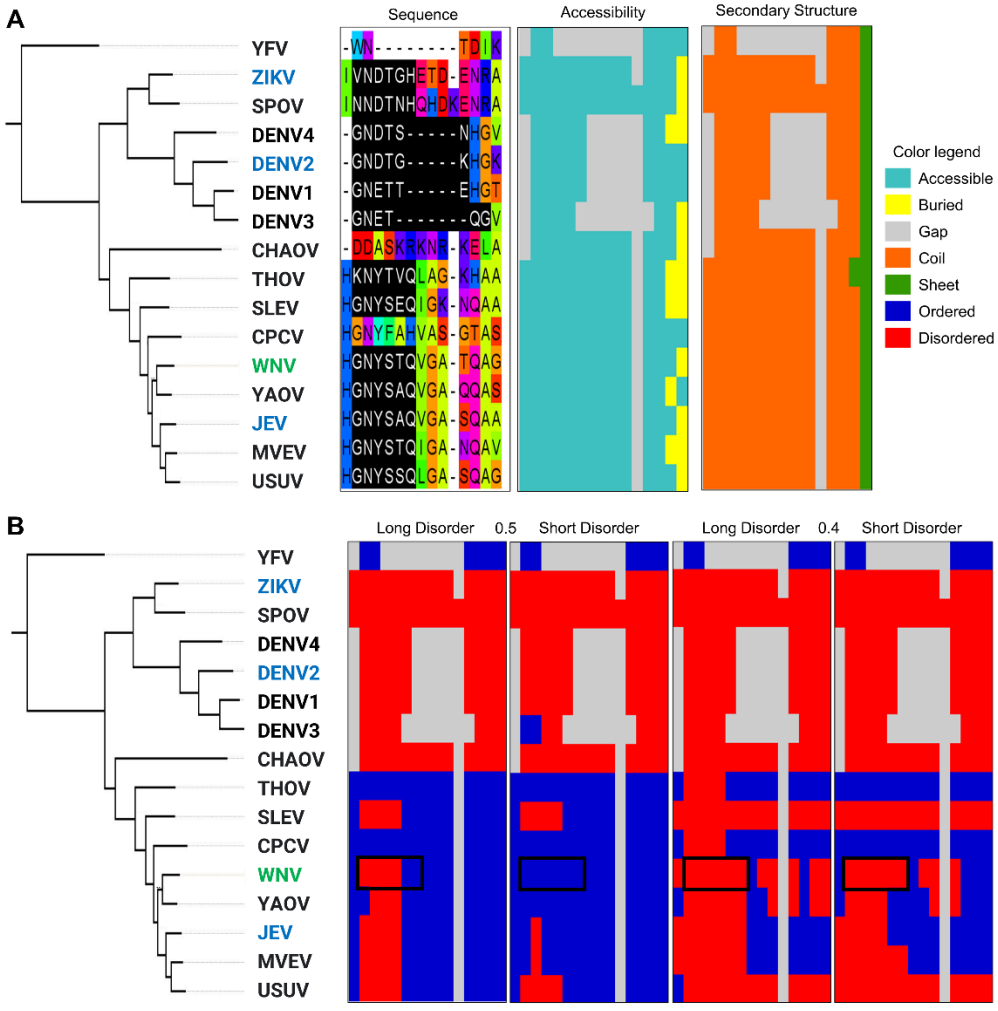

Figure 9. Phylogenetic tree of West Nile Virus (WNV) envelope protein illustrating the evolution of structural properties of a MOD_N-GLC_1 motif. The tree, rooted by the outgroup Yellow Fever virus (YFV)), shows WNV in green and Zika virus (ZIKV), Dengue virus 2 (DENV2), and Japanese Encephalitis Virus (JEV) that have been shown to be glycosylated in this position but that are not in the ELM database in blue. The tree is shown next to an excerpt from the multiple sequence alignment with the MOD_N-GLC_1 motif pattern highlighted in black, followed by the same alignment excerpt colored by the accessibility and secondary structure of the residues (A), and by disorder using both 0.5 and 0.4 cutoff values for long IUPRED2A and short IUPRED2A disorder, with the location of the WNV MOD_N-GLC_1 motif shown by the black box (B). For further details, see Figure S5.

LIG_Rb_LxCxE_1 is Less Disordered in Viruses

The LIG_Rb_LxCxE_1 motif is an amino acid sequence with a pattern shortly represented as LxCxE [48]. This motif is recognized by the tumor suppressors retinoblastoma protein (Rb), p107, and p130 involved in impeding G to S phase cell cycle progression [49]. Rb inhibits gene transcription through interactions with LIG_Rb_LxCxE_1 on the transcription factor E2F. Phosphorylation of the Rb protein initiates the release of E2F, and subsequently E2F-DNA binding activates cell cycle progression [48–51]. The LIG_Rb_LxCxE_1 motif has been found in viruses, especially DNA viruses [48,49,51–53]. Viral proteins displaying the LIG_Rb_LxCxE_1 motif bind to the Rb protein and leave the E2F transcription factor able to stimulate the cell cycle progression. Once cells replicate, the viruses take advantage of the replication enzymes to replicate their genome [54–56].

There were 32 SLiMs of the LIG_Rb_LxCxE_1 motif, 14 from viruses and 18 from eukaryotes. An analysis of the distribution of long and short IUPRED2A MIDS and MCCS values and long and short disorder and coil confidence per amino acid residue was performed. The LIG_Rb_LxCxE_1 motifs from viruses and eukaryotes show no significant difference in long and short MIDS, and short disorder score per residue values

(Figure 10), but long disorder score per residue, MCCS values, and coil confidence per residue are all higher for eukaryotic LIG_Rb_LxCxE_1 motifs (p-value = 1.88×10^{-3}, 3.30×10^{-3}, and 0, respectively) (Figure 10). The majority of the coil confidence score per residue was above 0.8 in eukaryotes. In contrast, viruses showed a wide distribution of coil confidence per residue. These results suggest differences in the binding mechanism between LIG_Rb_LxCxE_1 instances from some viral proteins vs. eukaryotic proteins and that the eukaryotic LIG_Rb_LxCxE_1 motif in eukaryotes are found in disordered domains that are composed of a long sequence of amino acids, while viral proteins are less disordered and found in short disordered sequence regions.

The variation in coil confidence observed for viruses indicates that some viral LIG_Rb_LxCxE_1 instances have more secondary structure content than others and could demonstrate different affinity for Rb. Eukaryotes rely on transient interactions with the retinoblastoma proteins and a dynamic regulation of the cell cycle process where high affinity would be detrimental. Some viruses may be similar to eukaryotes, while others may display more of the secondary structure needed for the interaction to occur, resulting in higher affinity. Unlike eukaryotes, viruses would benefit from blocking the retinoblastoma protein from inhibiting transcription factor E2F to ensure the progression of the cell cycle [56,57].

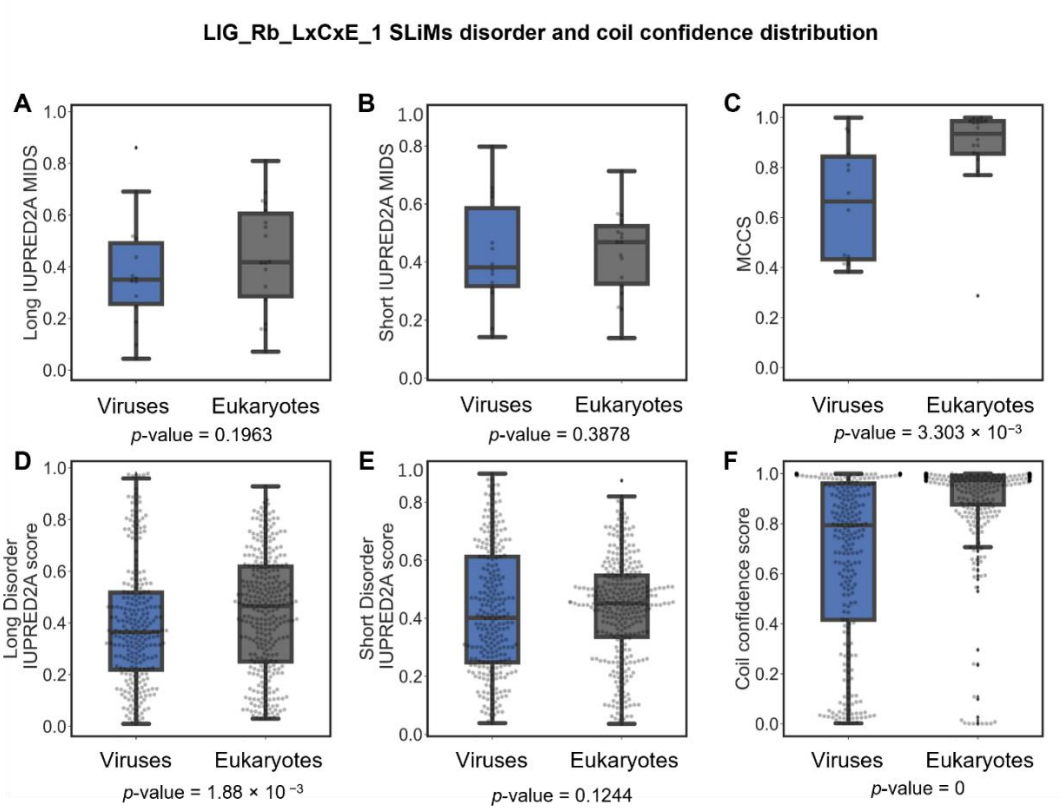

Figure 10. Disorder score and coil confidence distributions in viruses and eukaryotes for the LIG_Rb_LxCxE_1 motif. Boxplots and swarm plot distribution for SLiMs long IUPRED2A MIDS (A), short IUPRED2A MIDS (B), MCCS (C), individual long IUPRED2A disorder scores per residue for SLiMs (D), individual short IUPRED2A disorder scores per residue for SLiMs (E), and individual coil confidence scores per residue for SLiMs (F).

For the LxCxE motif, two instances from DNA viruses in the ELM database were annotated with a PDB structure; the large T antigen protein for Simian V40 virus (PDB ID: 1GH6 [58]) and E7 protein from human papillomavirus type 16 (PDB ID: 1GUX [59]). Neither structure represents the full-length proteins but instead truncated peptides of the LxCxE motif bound to human Rb. The secondary structure of the two motifs from the PDB structures based on DSSP reveals that the large T antigen protein is bound in a coil conformation, and the E7 protein is bound in a beta-strand conformation (Figure 11). Both motifs have high MCCS values of 0.94 and 0.95, respectively, and similar MIDS

values (0.50 and 0.51, respectively) for long IUPRED2A prediction and similarly, MIDS values (0.64 and 0.65, respectively) for short IUPRED2A prediction. Previous research has shown that the E7 motif showed a lower K_d value than the Simian V40 T antigen protein and higher binding affinity towards the Rb protein than native eukaryotic proteins (reviewed in [51]). The low variability in high MCCS values for the eukaryotic motifs suggests that its flexibility is vital for its function. Furthermore, the differences in coil confidence and binding affinity of the LxCxE motifs in eukaryotic and viral proteins indicate a selection for high coil confidence and against high-affinity binding for Rb protein to maintain the transient regulatory binding inside the cells for eukaryotes. For viral proteins, these motifs may at first occur by chance in a near-neutral manner, but subsequent amino acid substitutions may improve Rb-binding and increase selective pressure to improve the strength of the interaction.

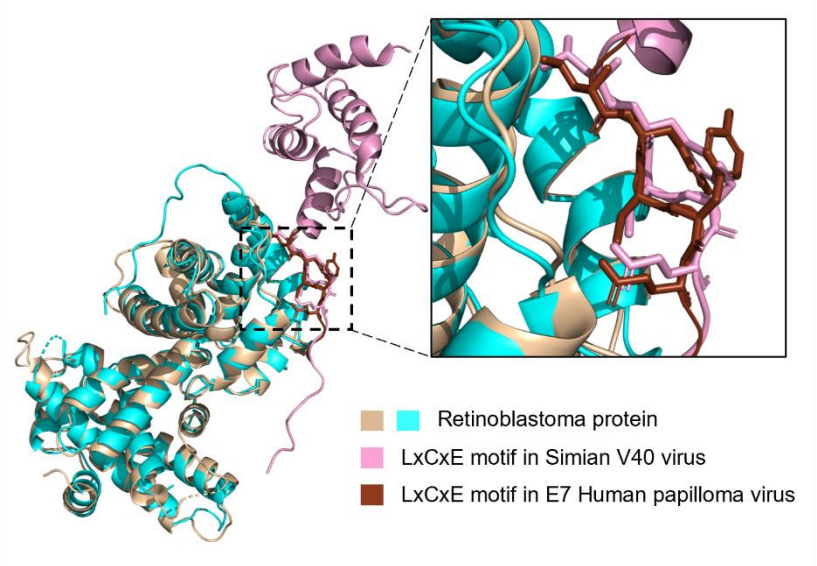

Figure 11. LIG_Rb_LxCxE_1 motif segment from Simian V40 (large T antigen protein) and Human papillomaviruses (E7) proteins in a bound state with retinoblastoma protein. The complete structures from PDB ID: 1GH6 and PDB ID: 1GUX are aligned, and a closer view of the LxCxE binding site is shown. Retinoblastoma

protein (beige and cyan) is rendered as a cartoon. Large T antigen protein is shown as cartoon (dark pink). The E7 of the Human papillomavirus motif segment is shown as ribbon (brown). The LxCxE motif in both proteins is shown as sticks. The structural alignment of the entire two structures was performed in PyMOL (PyMOL Molecular Graphics System, Version 4.6).

Conclusion

Based on the currently available data from the ELM database, we have explored the potential differences in sequence-based structural features between true positive SLiMs in different taxonomic groups: eukaryotes, bacteria, and viruses. We find that viral SLiMs often are less disordered than SLiMs from eukaryotes and bacteria, which seems to stem from different ELM functionality type compositions across taxonomic groups rather than differences in disorder for the equivalent SLiM. For the same SLiMs, the disorder content is in good agreement across taxonomic groups, but exceptions exist. Proteins harboring SLiMs are overall less disordered in viruses than in eukaryotes and bacteria, but for all taxonomic groups, a peak in disorder is observed for the SLiM containing region based on short IUPRED2A prediction. For long IUPRED2A prediction a small dip in disorder score is seen for SLiMs in eukaryotes as compared to the immediate flanking region but this dip is missing in viruses, but overall, the SLiMs containing region is more disordered than the rest of the protein. We find that most SLiMs across all taxonomic groups in our study are devoid of secondary structure and instead in a loop or coil conformation. We analyzed coil confidence and found that proteins harboring SLiMs peak in coil confidence at the SLiM. While proteins from viruses again have lower coil confidence overall, high coil confidence and coil content describe most SLiMs.

Disorder has been discussed as one of the most critical attributes in previous studies on SLiMs [5,25,60]. Our analysis of true positive SLiMs shows that classifying SLiMs as

false positives based on their lack of disorder is not feasible. Based on the experimentally verified SLiMs in this study, classifying SLiMs based on coil confidence would yield better results. However, no comparison of true positives vs. actual false positive SLiMs could be completed due to lack of such data. While the current study did not investigate the evolutionary dynamics of pathogenic SLiMs, such studies including 3D structural features, can bring further insights to molecular mimicry and host-pathogen interactions.

We have illuminated characteristics of SLiMs that may play a role in how pathogens utilize molecular mimicry of SLiMs to alter the host cell machinery to their advantage. We find that SLiMs from pathogens occasionally present vastly different structural characteristics than the same SLiM in the host. It is plausible that molecular mimicry is mediated through a more limited set of conformations than in the host, and different mechanisms of binding cannot be ruled out. However, the dataset of equivalent SLiMs from eukaryotes and their pathogens is limited and biased towards certain ELM types. Another limitation stems from an uneven distribution of SLiMs from closely related homologs. If the ELM database contains the same motif from the homologous conserved proteins, their characteristics can bias the results. When more data is available, this can be corrected for. An emphasis to experimentally verify more interactions involving SLiMs in pathogens is warranted to improve our understanding of molecular mimicry and host-pathogen interactions. In our analysis, we showed additional viruses that possess the same pattern for the MOD_N-GLC_1 motif from literature but not included in the ELM database. Unlike most MOD_N-GLC_1 entries from the ELM database, the additional instances were overall disordered. As more data becomes readily available, analyses and discoveries can be improved, and enhanced methods for identifying host-pathogen PPI

facilitated through molecular mimicry can be developed. Recent work by Wadie et al. applied structural and functional filters with information from viral SLiMs to enhance functional motif discovery in humans [61]. They used a low IUPRED2A disorder cutoff value of 0.2 to differentiate between functional and not functional SLiMs, but as we show here, caution must be taken when filtering viral SLiMs by IUPRED2A disorder even when using a very low cutoff value. A better understanding of the mechanistic differences displayed between the same SLiM in pathogens and their hosts holds promise for improving the utility of SLiMs as therapeutic drug targets.

Methods

4.1 The ELM Dataset

The complete dataset of SLiM instances in the ELM database was downloaded on October 10, 2021. SLiMs annotated as True Positives were kept for further analysis. The taxonomic IDs for the organisms were extracted using NCBI taxonomy [62], and the True Positive (TP) instances were categorized according to their taxonomy: eukaryotes (taxonomic ID 2759), bacteria (taxonomic ID 2), and viruses (taxonomic ID 10239) and ELM type. SLiMs from eukaryotes and viruses were further divided into taxonomic subcategories. All complete protein sequences that harbor a SLiM were downloaded from the ELM database and used to extract the amino acid sequence for each instance based on the ELM regular expression patterns. The complete sequences were also used to generate sequence-based structural predictions for eukaryotes and bacteria. For viruses, since the downloaded data from the ELM database included polyproteins and not the individual proteins that contain the motif, a custom script was used to extract all the viral protein sequence including the ones in a polyprotein based on Uniprot database chain annotation.

Some viral proteins did not have a chain annotation, these were manually examined and added to the viral dataset. For motifs that were found in-between two proteins based on UniProt chain annotation, the complete length of the two proteins were used. Three polyproteins that did not meet these criteria were excluded from the dataset.

4.2. Sequence-Based Structural Predictions

4.2.1. Intrinsic Disorder Prediction

Intrinsic disorder propensity for the full-length SLiM containing proteins downloaded from the ELM database was predicted using the IUPRED2A webserver using both the default settings (IUPRED2A long disorder) and the IUPRED2A short disorder [63]. The long disorder option searches for long segments of disordered regions in proteins, while short disorder option searches for short segments in proteins that may have disorder property located in interdomain linkers or within domains. For each SLiM instance, both long and short disorder scores for its amino acids were extracted and used to calculate the percent of disorder per instance and the Mean IUPRED2A Disorder Score (MIDS). The percent disorder per instance was calculated based on how many residues were above a given cutoff divided by the total number of residues in the instance multiplied by 100. Two cutoffs, 0.4 and 0.5, respectively, were used. MIDS was calculated as the average disorder score for all residues per instance.

4.2.2. Relative Solvent Accessibility and Secondary Structure Predictions

A local installation of NetSurfP 2.0 [64] was used to predict relative solvent accessibility and secondary structure for all full-length SLiM containing proteins downloaded from the ELM database. Predictions were run using the HHblits method from the HHSuite [65] and uniclust30_2017_04 database [66]. Relative solvent

accessibility was determined using a cutoff of 0.25. The coil or secondary structure assignment was considered based on the three-state prediction. For each instance, the percent of solvent-accessible and coil residues per instance were calculated as for the percent disorder described in 4.2.1 section. Coil confidence was extracted from the results and used to calculate the Mean Coil Confidence Score (MCCS) as for the average MIDS for all residues per instance.

4.3. Phylogenetic Tree Analysis

To build the West Nile Virus (WNV) envelope protein phylogenetic tree, a protein BLAST [67] was done to determine homologous proteins using the NCBI accession of the WNV envelope protein (YP_001527877) that contain the MOD_N-GLC-1 motif. Extracted homologs of the protein were aligned with MAFFT using the L-INS-i setting [68] in Jalview [69]. IQ-Tree [70] using the default settings of automatic selection of the substitution model, branch support analysis using the ultrafast bootstrap method with default settings, and SH-alrt branch test with 1000 replicates was used to generate the tree. The tree was rooted on the outgroup virus (Yellow fever virus). The phylogenetic tree and the multiple sequence alignment were used to inspect the variability in sequence conservation and map the disorder and secondary structure properties onto the alignment, to allow exploring the differences in these properties between different clades visually.

4.4. Statistical Analysis

Non-parametric statistical testing with Mann-Whitney was performed using a simplified Bonferroni multiple hypothesis testing correction (adjusted *p*-value = *p*-value multiplied by the number of tests, compared to alpha-value = 0.05) to infer statistically significant

differences between groups. Spearman correlation analysis was performed to test the correlation between groups. Both tests were performed using the SciPy module [71].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pathogens11050583/s1, Figure S1: Scatter plot for long disorder vs short disorder MIDS per instance; Figure S2: Scatter plot for eukaryotes long disorder vs short disorder MIDS per instance; Figure S3: Scatter plot for bacteria long disorder vs short disorder MIDS per instance; Figure S4: Scatter plot for viruses long disorder vs short disorder MIDS per instance; Figure S5: Phylogenetic tree of West Nile Virus envelope protein rooted by the outgroup Yellow Fever virus (YFV); Table S1: ELM dataset, prediction, and analysis; Table S2: SLiMs count and percentage per ELM type in eukaryotes, bacteria and viruses; Table S3: Disorder and coil confidence profile data for 100 residues around the SLiM; Table S4: Shared motifs prediction data between taxonomic groups

Data Availability Statement

Data, scripts, and Jupyter notebooks used for the analysis can be accessed at https://github.com/Heidy-Elkhaligy/Comparative-Analysis-of-Structural-Features-in-SLiMs-from-Eukaryotes-Bacteria-and-Viruses.git.

REFERENCES

1. Rao, V.S.; Srinivas, K.; Sujini, G.N.; Kumar, G.N.S. Protein-Protein Interaction Detection: Methods and Analysis. Int. J. Proteomics **2014**, 2014, 1–12, doi:10.1155/2014/147648.

2. Braun, P.; Gingras, A.C. History of protein–protein interactions: From egg-white to complex networks. Proteomics **2012**, 12, 1478–1498, doi:10.1002/PMIC.201100563.

3. Dunker, A.K.; Cortese, M.S.; Romero, P.; Iakoucheva, L.M.; Uversky, V.N. Flexible nets: The roles of intrinsic disorder in protein interaction networks. FEBS J. **2005**, 272, 5129–5148.

4. Uversky, V.N. The multifaceted roles of intrinsic disorder in protein complexes. FEBS Lett. **2015**, 589, 2498–2506, doi:10.1016/J.FEBSLET.2015.06.004.

5. Davey, N.E.; Van Roey, K.; Weatheritt, R.J.; Toedt, G.; Uyar, B.; Altenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T.J. Attributes of short linear motifs. Mol. Biosyst. **2012**, 8, 268–281, doi:10.1039/c1mb05231d.

6. Guven-Maiorov, E.; Tsai, C.J.; Nussinov, R. Pathogen mimicry of host protein-protein interfaces modulates immunity. Semin. Cell Dev. Biol. **2016**, 58, 136–145, doi:10.1016/J.SEMCDB.2016.06.004.

7. Sámano-Sánchez, H.; Gibson, T.J. Mimicry of Short Linear Motifs by Bacterial Pathogens: A Drugging Opportunity. Trends Biochem. Sci. **2020**, 45, 526–544, doi:10.1016/J.TIBS.2020.03.003.

8. Davey, N.E.; Travé, G.; Gibson, T.J. How viruses hijack cell regulation. Trends Biochem. Sci. **2011**, 36, 159–169, doi:10.1016/J.TIBS.2010.10.002.

9. Hraber, P.; O'Maille, P.E.; Silberfarb, A.; Davis-Anderson, K.; Generous, N.; McMahon, B.H.; Fair, J.M. Resources to Discover and Use Short Linear Motifs in Viral Proteins. Trends Biotechnol. **2020**, 38, 113–127.

10. Fuxreiter, M.; Tompa, P.; Simon, I. Local structural disorder imparts plasticity on linear motifs. Bioinformatics **2007**, 23, 950–956, doi:10.1093/bioinformatics/btm035.

11. Davey, N.E.; Cyert, M.S.; Moses, A.M. Short linear motifs - ex nihilo evolution of protein regulation. Cell Commun. Signal. **2015**, 13, doi:10.1186/S12964-015-0120-Z.

12. Uversky, V.N. Intrinsically disordered proteins and their “mysterious” (meta)physics. *Front. Phys.* **2019**, *7*, 10. https://doi.org/10.3389/FPHY.2019.00010.

13. Babu, M.M.; Kriwacki, R.W.; Pappu, R. V. Versatility from protein disorder. Science (80-.). **2012**, 337, 1460–1461.

14. Edwards, R.J.; Paulsen, K.; Aguilar Gomez, C.M.; Pérez-Bercoff, Å. Computational Prediction of Disordered Protein Motifs Using SLiMSuite. Methods Mol. Biol. **2020**, 2141, 37–72, doi:10.1007/978-1-0716-0524-0_3.

15. Gould, C.M.; Diella, F.; Via, A.; Puntervoll, P.; Gemünd, C.; Chabanis-Davidson, S.; Michael, S.; Sayadi, A.; Bryne, J.C.; Chica, C.; et al. ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res. **2010**, 38, D167–D180, doi:10.1093/NAR/GKP1016.

16. Gibson, T.J.; Dinkel, H.; Van Roey, K.; Diella, F. Experimental detection of short regulatory motifs in eukaryotic proteins: Tips for good practice as well as for bad. *Cell Commun. Signal.* **2015**, *13*, 42. https://doi.org/10.1186/S12964-015-0121-Y.

17. Dinkel, H.; Sticht, H. A computational strategy for the prediction of functional linear peptide motifs in proteins. Bioinformatics **2007**, 23, 3297–3303, doi:10.1093/bioinformatics/btm524.

18. Via, A.; Gould, C.M.; Gemünd, C.; Gibson, T.J.; Helmer-Citterich, M. A structure filter for the Eukaryotic Linear Motif Resource. BMC Bioinformatics **2009**, *10*, 351. https://doi.org/10.1186/1471-2105-10-351.

19. Davey, N.E.; Shields, D.C.; Edwards, R.J. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics **2009**, 25, 443–450, doi:10.1093/bioinformatics/btn664.

20. Elkhaligy, H.; Balbin, C.A.; Gonzalez, J.L.; Liberatore, T.; Siltberg-Liberles, J. Dynamic, but Not Necessarily Disordered, Human-Virus Interactions Mediated through SLiMs in Viral Proteins. Viruses **2021**, Vol. 13, Page 2369 2021, 13, 2369, doi:10.3390/V13122369.

21. Pushker, R.; Mooney, C.; Davey, N.E.; Jacqué, J.-M.; Shields, D.C. Marked variability in the extent of protein disorder within and between viral families. PLoS One **2013**, 8, e60724, doi:10.1371/journal.pone.0060724.

22. Kastano, K.; Erdős, G.; Mier, P.; Alanis-Lobato, G.; Promponas, V.J.; Dosztányi, Z.; Andrade-Navarro, M.A. Evolutionary Study of Disorder in Protein Sequences. Biomolecules **2020**, 10, 1413, doi:10.3390/BIOM10101413.

23. Peng, Z.; Yan, J.; Fan, X.; Mizianty, M.J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V.N.; Kurgan, L. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell. Mol. Life Sci. **2015**, 72, 137–151, doi:10.1007/S00018-014-1661-9.

24. Kumar, M.; Gouw, M.; Michael, S.; Sámano-Sánchez, H.; Pancsa, R.; Glavina, J.; Diakogianni, A.; Valverde, J.A.; Bukirova, D.; Čalyševa, J.; et al. ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res. **2020**, 48, D296–D306, doi:10.1093/NAR/GKZ1030.

25. Van Der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; et al. Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. **2014**, 114, 6589–6631, doi:10.1021/CR400525M.

26. Dyson, H.J.; Wright, P.E. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. **2005**, 6, 197–208.

27. Sharma, R.; Raduly, Z.; Miskei, M.; Fuxreiter, M. Fuzzy complexes: Specific binding without complete folding. FEBS Lett. **2015**, 589, 2533–2542, doi:10.1016/J.FEBSLET.2015.07.022.

28. Fuxreiter, M. Fuzziness: linking regulation to protein dynamics. Mol. Biosyst. **2011**, 8, 168–177, doi:10.1039/C1MB05234A.

29. Lis, H.; Sharon, N. Protein glycosylation: Structural and functional aspects. Eur. J. Biochem. **1993**, 218, 1–27.

30. Vigerust, D.J.; Shepherd, V.L. Virus glycosylation: role in virulence and immune interactions. Trends Microbiol. **2007**, 15, 211–218.

31. Bosques, C.J.; Tschampel, S.M.; Woods, R.J.; Imperiali, B. Effects of glycosylation on peptide conformation: A synergistic experimental and computational study. J. Am. Chem. Soc. **2004**, 126, 8421–8425, doi:10.1021/JA0496266/SUPPL_FILE/JA0496266SI20040121_063620.PDF.

32. Breitling, J.; Aebi, M. N-Linked Protein Glycosylation in the Endoplasmic Reticulum. Cold Spring Harb. Perspect. Biol. **2013**, 5, a013359. doi:10.1101/CSHPERSPECT.A013359.

33. Ruiz-Canada, C.; Kelleher, D.J.; Gilmore, R. Cotranslational and Posttranslational N-Glycosylation of Polypeptides by Distinct Mammalian OST Isoforms. Cell **2009**, 136, 272–283, doi:10.1016/J.CELL.2008.11.047.

34. Mohanty, S.; Chaudhary, B.P.; Zoetewey, D. Structural Insight into the Mechanism of N-Linked Glycosylation by Oligosaccharyltransferase. Biomolecules **2020**, *10*, 624. doi:10.3390/BIOM10040624.

35. Kelleher, D.J.; Karaoglu, D.; Mandon, E.C.; Gilmore, R. Oligosaccharyltransferase Isoforms that Contain Different Catalytic STT3 Subunits Have Distinct Enzymatic Properties. Mol. Cell **2003**, 12, 101–111, doi:10.1016/S1097-2765(03)00243-0.

36. Nybakken, G.E.; Nelson, C.A.; Chen, B.R.; Diamond, M.S.; Fremont, D.H. Crystal Structure of the West Nile Virus Envelope Glycoprotein. J. Virol. **2006**, 80, 11467, doi:10.1128/JVI.01125-06.

37. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **1983**, 22, 2577–2637, doi:10.1002/BIP.360221211.

38. Cherrier, M. V.; Kaufmann, B.; Nybakken, G.E.; Lok, S.M.; Warren, J.T.; Chen, B.R.; Nelson, C.A.; Kostyuchenko, V.A.; Holdaway, H.A.; Chipman, P.R.; et al. Structural basis for the preferential recognition of immature flaviviruses by a fusion-loop antibody. EMBO J. **2009**, 28, 3269, doi:10.1038/EMBOJ.2009.245.

39. Gall, T. Le; Romero, P.R.; Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Intrinsic disorder in the protein data bank. J. Biomol. Struct. Dyn. **2007**, 24, 325–341, doi:10.1080/07391102.2007.10507123.

40. Oldfield, C.J.; Xue, B.; Van, Y.Y.; Ulrich, E.L.; Markley, J.L.; Dunker, A.K.; Uversky, V.N. Utilization of protein intrinsic disorder knowledge in structural proteomics. Biochim. Biophys. Acta **2013**, 1834, 487, doi:10.1016/J.BBAPAP.2012.12.003.

41. Zhang, Y.; Stec, B.; Godzik, A. Between order and disorder in protein structures – analysis of "dual personality" fragments in proteins. Structure **2007**, 15, 1141, doi:10.1016/J.STR.2007.07.012.

42. Hanna, S.L.; Pierson, T.C.; Sanchez, M.D.; Ahmed, A.A.; Murtadha, M.M.; Doms, R.W. N-Linked Glycosylation of West Nile Virus Envelope Proteins Influences Particle Assembly and Infectivity. J. Virol. **2005**, 79, 13262, doi:10.1128/JVI.79.21.13262-13274.2005.

43. Mondotte, J.A.; Lozach, P.-Y.; Amara, A.; Gamarnik, A. V. Essential role of dengue virus envelope protein N glycosylation at asparagine-67 during viral propagation. J. Virol. **2007**, 81, 7136–7148, doi:10.1128/JVI.00116-07.

44. Carbaugh, D.L.; Baric, R.S.; Lazear, H.M. Envelope Protein Glycosylation Mediates Zika Virus Pathogenesis. J. Virol. **2019**, 93, e00113-19. doi:10.1128/JVI.00113-19.

45. Moudy, R.M.; Payne, A.F.; Dodson, B.L.; Kramer, L.D. Requirement of Glycosylation of West Nile Virus Envelope Protein for Infection of, but Not Spread within, Culex quinquefasciatus Mosquito Vectors. Am. J. Trop. Med. Hyg. **2011**, 85, 374, doi:10.4269/AJTMH.2011.10-0697.

46. Fall, G.; di Paola, N.; Faye, M.; Dia, M.; Freire, C.C.M.; Loucoubar, C.; Zanotto, P.M.A.; Faye, O.; Sall, A.A. Biological and phylogenetic characteristics of West African lineages of West Nile virus. *PLoS Negl. Trop. Dis.* **2017**, *11*, e0006078..

47. Wang, P.; Hu, K.; Luo, S.; Zhang, M.; Deng, X.; Li, C.; Jin, W.; Hu, B.; He, S.; Li, M.; et al. DC-SIGN as an attachment factor mediates Japanese encephalitis virus infection of human dendritic cells via interaction with a single high-mannose residue of viral E glycoprotein. Virology **2016**, 488, 108–119, doi:10.1016/J.VIROL.2015.11.006.

48. Dahiya, A.; Gavin, M.R.; Luo, R.X.; Dean, D.C. Role of the LXCXE Binding Site in Rb Function. Mol. Cell. Biol. **2000**, 20, 6799–6805, doi:10.1128/mcb.20.18.6799-6805.2000.

49. Burkhart, D.L.; Sage, J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. Nat. Rev. Cancer **2008**, 8, 671, doi:10.1038/NRC2399.

50. Fischer, M.; Müller, G.A. Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.* **2017**, 52, 638–662, doi:10.1080/10409238.2017.1360836.

51. Palopoli, N.; Foutel, N.S.G.; Gibson, T.J.; Chemes, L.B. Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity. Protein Eng. Des. Sel. **2018**, 31, 69–77, doi:10.1093/PROTEIN/GZX068.

52. Felsani, A.; Mileo, A.M.; Paggi, M.G. Retinoblastoma family proteins as key targets of the small DNA virus oncoproteins. Oncogene **2006**, 25, 5277–5285.

53. Narisawa-Saito, M.; Kiyono, T. Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: Roles of E6 and E7 proteins. Cancer Sci. **2007**, 98, 1505–1511.

54. Helt, A.M.; Galloway, D.A. Mechanisms by which DNA tumor virus oncoproteins target the Rb family of pocket proteins. Carcinogenesis **2003**, 24, 159–169, doi:10.1093/CARCIN/24.2.159.

55. Caracciolo, V.; Reiss, K.; Khalili, K.; De Falco, G.; Giordano, A. Role of the interaction between large T antigen and Rb family members in the oncogenicity of JC virus. Oncogene **2006**, 25, 5294–5301. doi:10.1038/sj.onc.1209681.

56. Fan, Y.; Sanyal, S.; Bruzzone, R. Breaking Bad: How Viruses Subvert the Cell Cycle. Front. Cell. Infect. Microbiol. **2018**, 8, 396, doi:10.3389/FCIMB.2018.00396/BIBTEX.

57. Chemes, L.B.; Sánchez, I.E.; De Prat-Gay, G. Kinetic recognition of the retinoblastoma tumor suppressor by a specific protein target. J. Mol. Biol. **2011**, 412, 267–284, doi:10.1016/J.JMB.2011.07.015.

58. Kim, H.Y.; Ahn, B.Y.; Cho, Y. Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen. EMBO J. **2001**, 20, 295, doi:10.1093/EMBOJ/20.1.295.

59. Lee, J.O.; Russo, A.A.; Pavletich, N.P. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. Nature **1998**, 391, 859–865, doi:10.1038/36038.

60. Davey, N.E.; Seo, M.-H.; Yadav, V.K.; Jeon, J.; Nim, S.; Krystkowiak, I.; Blikstad, C.; Dong, D.; Markova, N.; Kim, P.M.; et al. Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. FEBS J. **2017**, 284, 485–498, doi:10.1111/FEBS.13995.

61. Wadie, B.; Kleshchevnikov, V.; Sandaltzopoulou, E.; Benz, C.; Correspondence, E.P.; Petsalaki, E. Use of viral motif mimicry improves the proteome-wide discovery of human linear motifs. Cell Rep. **2022**, 39, 110764, doi:10.1016/J.CELREP.2022.110764.

62. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; McVeigh, R.; O'Neill, K.; Robbertse, B.; et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020, **2020**, baaa062. doi:10.1093/DATABASE/BAAA062.

63. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. **2018**, 46, W329–W337, doi:10.1093/nar/gky384.

64. Klausen, M.S.; Jespersen, M.C.; Nielsen, H.; Jensen, K.K.; Jurtz, V.I.; Sønderby, C.K.; Sommer, M.O.A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins Struct. Funct. Bioinforma. **2019**, 87, 520–527, doi:10.1002/prot.25674.

65. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods **2011**, *9*, 173–175, doi:10.1038/nmeth.1818.

66. Mirdita, M.; Von Den Driesch, L.; Galiez, C.; Martin, M.J.; Soding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. **2017**, 45, D170–D176, doi:10.1093/NAR/GKW1081.

67. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. **1990**, 215, 403–410, doi:10.1016/S0022-2836(05)80360-2.

68. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **2013**, 30, 772–80, doi:10.1093/molbev/mst010.

69. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics **2009**, 25, 1189–1191, doi:10.1093/bioinformatics/btp033.

70. Nguyen, L.T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. **2015**, 32, 268, doi:10.1093/MOLBEV/MSU300.

71. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods **2020**, 17, 261–272, doi:10.1038/s41592-019-0686-2.

SUMMARY AND FUTURE DIRECTIONS

In this thesis, we highlighted the importance of studying the properties of Short Linear Motifs (SLiMs). In the first chapter, the introduction, we reviewed some of the previously published articles that presented the importance of intrinsically disordered protein regions that harbors SLiMs in promoting protein-protein interactions inside our cells. SLiMs can modulate and regulate cellular processes, especially in eukaryotes. The characteristics of eukaryotic SLiMs identified thus far, such as being in disordered regions or surface accessibility or being in loops in the protein structure, are essential for filtering true positive eukaryotic SLiMs. We also shed light on the molecular mimicry mechanism utilized by pathogens through mimicking the host SLiMs and how that may be vital for the fitness and pathogenicity of the infecting virus or bacteria. The literature review illuminated some gaps that were addressed in this thesis, for instance, the absence of a comprehensive study on the attributes of SLiMs from pathogens, such as bacteria and viruses. In addition, the lack of a comparative study that compares the differences or similarities between the sequence-based structure properties of SLiMs between different taxonomic groups and different ELM types.

The second chapter explored the true positive viral SLiMs in the ELM database to unravel their structural properties using a sequence-based structural approach. A multiple sequence alignment of selected viral instances was used to explore their conservation and sequence-based structural differences among homologous viral proteins. The analysis revealed that viral SLiMs are not found in disordered regions, contradicting the notion established before about the eukaryotic motifs. In addition, they are mainly found in an accessible and coil region. Moreover, the analysis of the selected instances in different phases of the viral life cycle revealed that not all residues in the instances are in a

conserved protein region. With the recent Covid-19 pandemic and the perplexing variations between patients, more attention should be given to discovering the functional vs. non-functional SLiMs. Uncovering SLiMs features may help locate functional SLiMs, not only in SARS-CoV-2 virus, but also in other emerging pathogens.

In chapter 3, a comparative analysis to investigate the true positive SLiMs in the ELM database to detect any sequence-based structural differences between three taxonomic groups (eukaryotes, bacteria, and viruses) and between ELM types was presented. Previously, disorder was analyzed as a mean disorder score per instance. Following the same concept, we created a new metric that can enable us to measure the coil properties of the SLiMs instances by using the mean coil confidence score per instance. We noticed that pathogens seem to have a similar trend as eukaryotes, where the majority are surface accessible, based on percent accessibility, and found in coil regions based on both percent coil and the mean coil confidence score. Disorder analysis revealed that although proteomes from eukaryotes tend to have more disorder content overall than proteomes from bacteria and viruses, SLiMs from eukaryotes and bacteria were more disordered and had higher coil confidence than viral SLiMs.

When analyzing SLiMs by ELM type, the modification motifs of pathogens demonstrated a low disorder content. We further compared the SLiMs in pathogens with their eukaryotic counterparts, pathogens sequence-based structure properties such as MIDS and MCCS values tend to correlate moderately with eukaryotes. To delve deeper into the differences and/or similarities between pathogens and their counterpart eukaryotic SLiMs, we only explored two examples due to the low number of pathogenic instances. The MOD_N-GLC_1, the glycosylation motif, and the LIG_Rb_LxCxE_1, the

motif responsible for interaction with the retinoblastoma protein. The former motif showed no significant difference in the disorder and coil confidence properties between viruses and eukaryotes. However, when analyzing the only instance with an annotated structure in the ELM database, West Nile Virus envelope glycosylation motif, we found that the primary residue (Arginine) in the motif is found in a nearly conserved site in other viral envelope proteins in related homologs, and some of them harbor the same functional motif based on literature search. The functional motif in the related homologs were all in an accessible coil region. However, the disorder varies depending on the type and cutoff value used to determine the disorder, indicating that a better understanding of the differences or similarities in this motif can be achieved when more data is available. For the latter, retinoblastoma protein-binding motif, our analysis revealed a significant difference between the eukaryotes and viruses using the long disorder MIDS, and MCCS values, where eukaryotes are significantly higher than viruses. Such significance implies that the eukaryotes genome evolves under the multifaceted constraint that differs from the constraint acting on viruses. Thus, disorder and being in a coil is advantageous at binding interfaces that rely on conformational transitions where SLiMs may act as molecular on/off switches. Nevertheless, disorder and being unstructured (coil) may become less advantageous when an ordered viral SLiM mimics a functional conformation of a host SLiM so that it is always switched on or off.

The work done in this thesis emphasized the existence of variations as well as resemblances between eukaryotic and pathogenic SLiMs, which demands additional experimental exploration to enhance the computational detection of true positive SLiMs and functional pathogenic SLiMs rapidly and effortlessly. In chapters 2 and 3, we highlighted the current filtration methods' limitations in identifying functional SLiMs in all taxonomic groups. Moreover, examples of various viruses' multiple sequence alignment and phylogenetic analysis displayed the presence of clade- or virus-specific patterns for some of the motifs that may have a role in the pathogenicity differences exhibited between related viruses or even different strains of the same virus. For instance, in chapter 2, the literature showed that the PDZ-binding domain in some strains of the E7 protein of HPV type 16 and 18 makes the host cells infected with these viruses lose cellular polarity and increase the metastatic incidence of cancer caused by that viral infection. The E7 homologous protein's multiple sequence alignment showed that the motif is present in other strains. However, it is located in a highly variable region in the alignment, and other HPV strains show differences in their disorder content, which may imply that the function of this motif is sequence and structural dependent. Additionally, in chapter 3, the phylogenetic analysis of the glycosylation motif of flaviviruses envelope protein displayed a highly conserved arginine residue at site 154 in the multiple sequence alignment. Previous experimental studies revealed that the loss of this specific glycosylation motif in some flaviviruses leads to the presence of a less infective species. This finding is crucial as it denotes that during viruses' evolution, the loss or gain of specific SLiMs may affect the pathogenicity and virulence of the newly emerging

viruses. Recently, another study demonstrated that viral SLiMs may be used to identify functional human SLiMs with higher accuracy.

Altogether, it implicates the importance of identifying and studying the pathogenic SLiMs, not only to be able to rapidly find a solution to conquer the current emerging pathogens but also to be prepared for any upcoming pandemics caused by infective pathogens.