Computational Analysis of the Linear Motif Mediated Subversion of the Human Protein Synthesis Machinery

Andrés Becerra, Victor A. Bucheli, Pedro A. Moreno,
School of Systems Engineering and Computer Science
Faculty of Engineering
Universidad del Valle
Email: {andres.becerra.sandoval, victor.bucheli, pedro.moreno}@correounivalle.edu.co

***Abstract*—We argue that virus-host interactions mediated by short linear motifs can be used to analyze common viral attack strategies. In this direction we develop a method for predicting interactions between human protein-synthesis machinery and viral proteins mediated by linear motifs in order to study common protein-synthesis subversion strategies. The method consists in finding viral instances of host linear motifs. We filter these instances by conservation in viral sequences, location in protein disordered regions and scarcity in randomized protein sets. With the filtered motifs we deduce virus-host interactions using the motif-domain associations in the Eukaryotic Linear Motifs (ELM) database. We validate the results against the Linear Motif mediated Protein Interaction Database (LMPID) and obtain a network of interactions between the human protein-synthesis machinery proteins and viruses influenza AH1N1, Dengue1, Ebola, MERS, Rotavirus, WestNile, and Zika.**

***Index Terms*—virus; host; protein; interaction; short; linear; motif; prediction; eukarya; protein-synthesis; subversion**

I. Introduction

The objective of this paper is to present a work in progress for predicting virus-host protein-protein interactions (VHPPIs) between several viruses and the human protein-synthesis machinery (HPSM) mediated by short linear motifs (SLiMs). Our motivation to conduct this study is to unveil common viral strategies to subvert protein translation.

There is no known virus that encodes a complete protein-synthesis system. This implies that viruses are forced to use the HPSM to translate their messenger RNA (mRNA) into products: microRNA (miRNA), peptides and proteins. Viruses must control the HPSM and disrupt innate host defense systems capable of disabling protein synthesis [1].

The control and disruption of host signaling pathways is conducted through VHPPIs like the ones DNA viruses engage with the PI3K–Akt–mTOR pathway (phosphatidylinositol 3-kinase-Akt-mammalian target of rapamycin) [2]. The consequences of VHPPIs can be as significant as the shutdown of host protein synthesis done by Rotavirus protein NSP3 [3].

There are open questions about the viral control of the HPSM like the role of phosphorylation in activity of protein eIF4E and how viral mRNA is preferentially translated [4]. These questions could be investigated with a systems biology approach.

Systems biology uses VHPPIs for the discovery of infection mechanisms [5]. However, the scarcity of virus-host PPIs with experimental evidence is an obstacle to system approaches [6]. This lack of data has encouraged the development of VHPPI prediction methods.

VHPPI prediction methods have been mostly based on machine learning classifiers like random forests [7] and support vector machines [8]–[10]. Most of these classifiers use protein sequences and other features like gene ontology (GO) function and gene expression as inputs to infer the interactions because structural data for viral proteins is scarce [11].

There are other prediction methods like information integration [12], asking experts [13], literature mining [14] and focusing on PPIs mediated by SLiMs [15].

We focus our study on SLiM-mediated interactions. The inference of this kind of interactions is guided by biological hypotheses like the conservation of motifs and localization of motifs in protein disordered regions.

Recently, the role of SLiMs has been studied in a wide set of viruses. These pathogens use SLiMs extensively as means to interact with host proteins [16]. Human proteins targeted by viruses have a high number of SLiMs [17].

If virus-host PPIs are divided in domain-motif interactions (DMI) and domain-domain interactions (DDI), DMI are the predominant ones. Furthermore, DMI are used by several viruses while DDI are virus-specific [17]. This supports our use of SLiMs as a way to find common viral subversion strategies.

Eukaryotic organisms use SLiM instances as as mechanisms to tune the regulation of multi-protein complexes. These instances are short, allowing viruses to evolve them de novo and retain them if they are useful to disrupt o subvert a host protein complex [18]. If the SLiM instances are encoded in different host genomic locations, the viral evolution of SLiM instances is robust in a virus-host coevolutionary arms race [19].

SLiMs are represented computationally as regular expressions like **PxIxIT** for the PCNA-binding PIP box motif of Flap endonuclease 1 (FEN1), where the **x** stands for any amino acid. A SLiM instance is a subsequence in a protein that

 ISBN: 978-1-61208-488-6

matches the regular expression, like **PRIEIT** in the human protein NFATC1 [18].

Viral instances of regular expressions representing host SLiMs can be found by chance. For this reason, filtering methods of viral instances must be implemented.

Evans et al. find that HIV-1 instances of human SLiMs are significantly conserved in HIV-1 proteins [15]. They propose a criterion to filter SLiMs if they are conserved above a 70% in the available viral sequences.

Hagai et al. propose two criteria to filter SLiMs: the first is based on SLiM location in protein disordered regions and the second in SLiM rarity in a big set of randomized (chimeric) proteins [16]. A SLiM is judged as rare, or hard to form by pure chance, if it is counted in less than a fraction of the sequences in the set of randomized proteins, e.g. 1% of the sequences.

We implement a combinaion of filtering criteria: 1) conservation, 2) location in disordered region and 3) difficulty to find the SLiM by chance. Our contribution is computational, the development of a platform to predict SLiM-mediated interactions that can be generalized to other subsystems and hosts. The clear limitation of our platform is our reliance on the ELM motifs database that makes the method appropriate for eukaryotic hosts only.

The organization of this paper is as follows. In Section III we present the results or our work. In Section II we describe the computational methods used and the Section IV contains the conclusion and directions for further research.

II. METHODS

Algorithmically, the prediction of SLiM-mediated VHPPI we propose is divided into: 1) collecting regular expressions representing SLiMs in the HPSM proteins, 2) finding instances of the collected SLiMs in viral proteins, 3) filtering the instances, 4) infer VHPPIs using SLiM instances in viral proteins and counter domains (CDs) in host proteins.

In order to complete the phases enumerated above we: 1) use the ELM database as a catalog of SLiMs [22], 2) implement software to find SLiM instances in protein sequences, 3) develop three filtering criteria described below, and 4) use the SLiM-domain associations in the ELM database together with Pfam protein-domain associations to infer protein-protein interactions [23].

A. Sequences and disorder prediction

HPSM proteins are taken from reference [1] and the Ribosomal Protein Gene database (RPG) [24]. All proteins are mapped to Uniprot identifiers in order to match protein entries in the ELM database [25].

Viruses are selected for their availability of protein sequences in the National Center for Biotechnology Information (NCBI) viral genomes resource: Dengue virus, West Nile virus, Middle East Respiratory Syndrome coronavirus (MERS), Ebolavirus, Rotavirus and Zika virus [26]. For influenza we choose type A, subtype H1N1, for Dengue we choose type 1, for Ebola the Zaire species.

We download every viral protein for each virus. For all viruses, we set the parameter region as any, the parameter "Full-length sequences only" to true and the parameter host as human. For Influenza AH1N1 proteins we set the parameter collapsed sequences, with the exception of proteins M1,M2 and NS2 for which the collapsed sequences option was deactivated.

For viruses Dengue type 1, West Nile and Zika the NCBI viral genomes resource gives the complete polyprotein sequence that must be manually cleaved. The viral reference genomes stored in Genbank files are computationally translated to protein sequences that are used as reference for cleaving the polyprotein into viral proteins.

Disorder prediction is computed with IUPred [27]. We develop a wrapper to call IUPred on each protein sequence to compute the disordered regions with a sliding-window algorithm proposed by Hagai et al. [16].

B. SLiMs

We download all the SLiMs, instances and interactions from the ELM database and create a SLiM dictionary indexed by the ELM unique identifiers containing the SLiM name, class and its full regular expression [28]. We develop scripts to compute for a set of sequences: the number of sequences with a given SLiM, the number of SLiM instances per protein, the number of SLiMs conserved above a percentage of sequences (set C) and the number of SLiMs in disordered regions (set D).

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

Finally, we use the scripts to generate the sets C, D and R for every viral protein using all the SLiMs in the ELM database.

C. Interactions

We compute the SLiM instances in viral proteins for all the human SLiM regular expressions in the set $C \cup D \cup R$. With the SLiM instances we infer PPIs between humans and the corresponding virus using the SLiM-domain associations in the ELM database and the protein-domain associations in the Pfam database [23]. We validate the interactions obtained with the LMPID database [29].

D. Analysis of the interactions

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

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

 ISBN: 978-1-61208-488-6

TABLE I
NUMBER OF INTERACTIONS PREDICTED WITH VIRAL PROTEINS

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

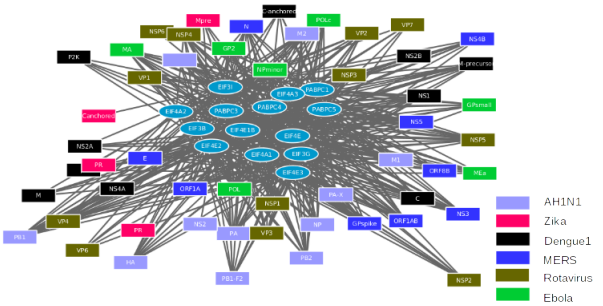

Fig. 1. Protein-protein interaction network predicted for protein-synthesis and viral proteins. Human protein-synthesis proteins are represented as ellipses and viral proteins as boxes. Boxes are colored differently for each virus.

The disrupting or wiring interactions are contrasted with the information in the KEGG pathway database [21] and gene ontology [20].

III. RESULTS

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

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

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

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

We obtain a network of interactions between human proteins in the HPSM subsystem and the proteins of the selected viruses represented in Figure 1.

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

TABLE II
DEGREE DISTRIBUTION FOR HUMAN PROTEINS

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

TABLE III
DEGREE DISTRIBUTION FOR VIRAL PROTEINS

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

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

We find that the protein EIF3I, the Eukaryotic translation initiation factor 3 subunit I is the hub of the HPSM system, with 78 interactions. EIF3I is involved in the formation of translation preinitiation complex, regulation of translational

TABLE IV
VIRAL PROTEINS WITH ONE INTERACTION (POTENTIALLY DISRUPTING)

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

 ISBN: 978-1-61208-488-6

TABLE V
VIRAL HUB PROTEINS (POTENTIALLY BRIDGING)

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

initiation and assembly of the eukaryotic 48S preinitiation complex [20]. The EIF3I protein is in the hsa03013 RNA transport KEGG pathway, in which it is part of a multifactor complex with EIF1, EIF2 and EIF5 [21].

We tried to validate the interactions found against the LMPID database but did not found any candidate interaction there. Perhaps the coverage of SLiM-mediated VHPPIs is too limited at the moment.

IV. CONCLUSION AND FUTURE WORK

We propose the prediction of SLiM-mediated host-virus PPIs between the human HPSM and some selected viruses. Further analysis of the interactions obtained might yield clues about common viral strategies for subverting protein translation.

Our main contribution is the combination of SLiM filtering methods. Having a general implementation of SLiM finding and filtering allows that the methods can be extended to other subsystems like the interferon [30] and apoptosis proteins [31] to investigate viral infection mechanisms at different stages. The methods can even be used with non-human eukaryotic hosts.

ACKNOWLEDGEMENTS

This study was supported by a Colciencias scholarship granted to AB. The authors would like to thank Dr. Aydin Tozeren for receiving AB in the Biomedical Engineering Department at Drexel university and suggesting the protein translation machinery as a subsystem to study.

REFERENCES

[1] D. Walsh and I. Mohr, "Viral subversion of the host protein synthesis machinery," *Nat. Rev. Microbiol.*, vol. 9, pp. 860–875, Dec 2011.

[2] N. J. Buchkovich, Y. Yu, C. A. Zampieri, and J. C. Alwine, "The TORrid affairs of viruses: effects of mammalian DNA viruses on the PI3K-Akt-mTOR signalling pathway," *Nat. Rev. Microbiol.*, vol. 6, pp. 266–275, Apr 2008.

[3] L. Padilla-Noriega, O. Paniagua, and S. Guzman-Leon, "Rotavirus protein NSP3 shuts off host cell protein synthesis," *Virology*, vol. 298, pp. 1–7, Jun 2002.

[4] S. Flint, V. Racaniello, G. Rall, and A. M. Skalka *Principles of Virology*. Third edition. American Society for Microbiology, 2009.

[5] S. Durmuş, T. Çakir, A. Özgür, and R. Guthke, "A review on computational systems biology of pathogen-host interactions," *Front Microbiol*, vol. 6, p. 235, 2015.

[6] S. D. Durmuş Tekir and K. O. Ülgen, "Systems biology of pathogen-host interaction: networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era," *Biotechnol J*, vol. 8, pp. 85–96, Jan 2013.

[7] S. Wuchty, "Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens," *PLoS ONE*, vol. 6, no. 11, p. e26960, 2011.

[8] M. D. Dyer, T. M. Murali, and B. W. Sobral, "Supervised learning and prediction of physical interactions between human and HIV proteins," *Infect. Genet. Evol.*, vol. 11, pp. 917–923, Jul 2011.

[9] G. Cui, C. Fang, and K. Han, "Prediction of protein-protein interactions between viruses and human by an SVM model," *BMC Bioinformatics*, vol. 13 Suppl 7, p. S5, 2012.

[10] R. K. Barman, S. Saha, and S. Das, "Prediction of interactions between viral and host proteins using supervised machine learning methods," *PLoS ONE*, vol. 9, no. 11, p. e112034, 2014.

[11] J. M. Doolittle and S. M. Gomez, "Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens," *Virol. J.*, vol. 7, p. 82, 2010.

[12] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman, "Prediction of interactions between HIV-1 and human proteins by information integration," *Pac Symp Biocomput*, pp. 516–527, 2009.

[13] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman, "Refining literature curated protein interactions using expert opinions," *Pac Symp Biocomput*, pp. 318–329, 2015.

[14] T. Thieu, S. Joshi, S. Warren, and D. Korkin, "Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches," *Bioinformatics*, vol. 28, pp. 867–875, Mar 2012.

[15] P. Evans, W. Dampier, L. Ungar, and A. Tozeren, "Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs," *BMC Med Genomics*, vol. 2, p. 27, 2009.

[16] T. Hagai, A. Azia, M. M. Babu, and R. Andino, "Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions," *Cell Rep*, vol. 7, pp. 1729–1739, Jun 2014.

[17] R. R. Halehalli and H. A. Nagarajaram, "Molecular principles of human virus protein-protein interactions," *Bioinformatics*, vol. 31, pp. 1025–1033, Apr 2015.

[18] N. E. Davey, M. S. Cyert, and A. M. Moses, "Short linear motifs - ex nihilo evolution of protein regulation," *Cell Commun. Signal*, vol. 13, no. 1, p. 43, 2015.

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

[20] M. Ashburner et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, vol. 25, pp. 25–29, May 2000.

[21] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Res.*, vol. 40, pp. D109–114, Jan 2012.

[22] H. Dinkel et al., "ELM–the database of eukaryotic linear motifs," *Nucleic Acids Res.*, vol. 40, pp. D242–251, Jan 2012.

[23] R. D. Finn et al., "The Pfam protein families database: towards a more sustainable future," *Nucleic Acids Res.*, vol. 44, pp. D279–285, Jan 2016.

[24] A. Nakao, M. Yoshihama, and N. Kenmochi, "RPG: the Ribosomal Protein Gene database," *Nucleic Acids Res.*, vol. 32, pp. D168–170, Jan 2004.

 ISBN: 978-1-61208-488-6

[25] A. Bateman et al, "UniProt: a hub for protein information," *Nucleic Acids Res.*, vol. 43, pp. D204–212, Jan 2015.
[26] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova, "NCBI viral genomes resource," *Nucleic Acids Res.*, vol. 43, pp. D571–577, Jan 2015.
[27] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, "The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins," *J. Mol. Biol.*, vol. 347, pp. 827–839, Apr 2005.
[28] H. Dinkel et al., "The eukaryotic linear motif resource ELM: 10 years and counting," *Nucleic Acids Res.*, vol. 42, pp. D259–266, Jan 2014.
[29] D. Sarkar, T. Jana, and S. Saha, "LMPID: a manually curated database of linear motifs mediating protein-protein interactions," *Database (Oxford)*, vol. 2015, 2015.
[30] V. Navratil, B. de Chassey, L. Meyniel, F. Pradezynski, P. Andre, C. Rabourdin-Combe, and V. Lotteau, "System-level comparison of protein-protein interactions between viruses and the human type I interferon system network," *J. Proteome Res.*, vol. 9, pp. 3527–3536, Jul 2010.
[31] S. E. Hasnain et al., "Host-pathogen interactions during apoptosis," *J. Biosci.*, vol. 28, pp. 349–358, Apr 2003.

 ISBN: 978-1-61208-488-6