*mimic*INT: a workflow for microbe-host protein interaction inference

Sébastien A. Choteau[1], Marceau Cristianini[1], Kevin Maldonado[1], Lilian Drets[1], Mégane Boujeant[1], Christine Brun[1,2], Lionel Spinelli[1,*], Andreas Zanzoni[1,*, #]

[1]Aix-Marseille Univ, INSERM, TAGC, UMR_S1090, Turing Centre for Living Systems, Marseille, France, [2]CNRS, Marseille, France

* Equal contribution.

To whom correspondence should be addressed: andreas.zanzoni@univ-amu.fr.

Abstract

The increasing incidence of emerging infectious diseases is posing serious global threats. Therefore, there is a clear need for developing computational methods that can assist and speed-up experimental research to better characterize the molecular mechanisms of microbial infections. In this context, we developed *mimic*INT, a freely available computational workflow for large-scale protein-protein interaction inference between microbe and human by detecting putative molecular mimicry elements that can mediate the interaction with host proteins: short linear motifs (SLiMs) and host-like globular domains. *mimic*INT exploits these putative elements to infer the interaction with human proteins by using known templates of domain-domain and SLiM-domain interaction templates. *mimic*INT provides *(i)* robust Monte-Carlo simulations to assess the statistical significance of SLiM detection which suffers from

false positive, and *(ii)* interaction specificity filter to account for differences between motif-binding domains of the same family.

*mimic*INT is implemented in Python and R, and it is available at: https://github.com/TAGC-NetworkBiology/mimicINT.

Introduction

Most pathogens interact with their hosts to reach an advantageous niche and ensure their successful dissemination. For instance, viruses interfere with important host-cell processes through protein-protein interactions to coordinate their life cycle (Yamauchi and Helenius, 2013). It has been shown that host cell networks subversion by pathogen proteins can be achieved through interface mimicry of endogenous interactions (i.e., interaction between host proteins) (Franzosa and Xia, 2011; Garamszegi *et al.*, 2013). This strategy relies on the presence in pathogen protein sequences of host-like elements, such as globular domains and short linear motifs (SLiMs), that can mediate the interaction with host proteins (Davey *et al.*, 2011; Hagai *et al.*, 2014; Via *et al.*, 2015).

Over the last years, many computational methods have been developed to predict pathogen-host protein interactions, some of which are based on the detection of sequence or structural mimicry elements (Arnold *et al.*, 2012; Nourani *et al.*, 2015). Such approaches allowed, for instance, to suggest potential molecular mechanisms underlying the implication of gastrointestinal bacteria in human cancer (Zanzoni *et al.*, 2017; Guven-Maiorov *et al.*, 2017) or to discriminate between viral strains with different oncogenic potential (Lasso *et al.*, 2019), thus showing that protein-protein interaction predictions can be instrumental in untangling microbe-host disease

associations. Nevertheless, the source code of many of these tools are not freely available to the community (e.g., (Becerra *et al.*, 2017; Guven-Maiorov *et al.*, 2017; Lasso *et al.*, 2019)) providing the predictions through a database (e.g., (Lasso *et al.*, 2019)), or can be only used through a web interface (e.g. (Guven-Maiorov *et al.*, 2020)), thus limiting the prediction reproducibility and tool usability.

In this context, and inspired by our previous work (Zanzoni *et al.*, 2017), we present *mimic*INT, a computational workflow for large-scale interaction inference between microbe and human proteins by detecting host-like elements and using experimentally identified interaction templates (Mosca *et al.*, 2014; Kumar *et al.*, 2020).

Implementation

*mimic*INT detects putative molecular mimicry elements in microbe sequences of interest that can mediate the interaction with host proteins (Figure 1). *mimic*INT is written in Python and R languages and exploits the Snakemake workflow manager for automated execution (Köster and Rahmann, 2018). It consists of four main steps: *(i)* the detection of host-like elements in microbe sequences; *(ii)* the collection of domains on the host protein *(iii)*; the interaction inferences between microbe and host proteins; and *(iv)* the functional enrichment analysis on the list of inferred host interactors.

In the first step, *mimic*INT takes as input the FASTA-formatted sequences of microbe proteins (*e.g.*, viral or other pathogen proteins susceptible to be found at the pathogen-host interface) to detect host-like elements: domains and SLiMs. The domain identification is performed by the stand-alone version of InterProScan (Jones

et al., 2014) using the domain signatures from the InterPro database (Blum *et al.*, 2021). By default, *mimic*INT retains InterProScan matches with an *E*-value below 10^{-5}, a threshold value commonly used for detecting profile-based domain signatures in protein sequences in the context of interaction inference (Schleker *et al.*, 2012). The host-like SLiMs detection exploits the motif definitions available in the ELM database (Kumar *et al.*, 2020) and is carried out by the SLiMProb tool from the SLiMSuite software package (Edwards *et al.*, 2020). As SLiMs are usually located in disordered regions (Davey *et al.*, 2012), SLiMProb uses the IUPred algorithm (Dosztányi, 2018) to compute the disorder propensity of each amino acid in the query sequences, and generates an average disorder propensity score for every detected SLiM occurrence. For SLiM detection, the default IUPred disorder propensity threshold is set to 0.2, a value commonly used to limit false negatives (Edwards and Palopoli, 2015; Edwards *et al.*, 2020), and the minimum size of the predicted disorder region is set to 5, the optimal size to detect true positive SLiM occurrences (Paulsen, 2019). Nevertheless, the user can choose all running parameters for the host-like element detection in the *mimic*INT configuration file.

In the second step, *mimic*INT gathered the domain annotations of the host proteins from the InterPro database through a REST API query.

In the third step, *mimic*INT infers the interactions between host and microbe proteins. This analysis takes as input the list of known interactions templates gathered from two resources: *(i)* the 3did database (Mosca *et al.*, 2014), a collection of domain-domain interactions extracted from three-dimensional protein structures (Rose *et al.*, 2013), and *(ii)* the ELM database (Kumar *et al.*, 2020) that provides a list of experimentally identified SLIM-domain interactions in Eukaryotes. The inference

checks whether any of the microbe proteins contains at least one domain or SLIM for which an interaction template is available. In this case, it infers the interaction between the given protein and all the host proteins containing the cognate domain (*i.e.*, the interacting domain in the template). As motif-binding domains of the same group, like SH3 or PDZ, show different interaction specificities (Gfeller *et al.*, 2011) for the SLiM-domain interaction inference, we have implemented a previously proposed strategy (Weatheritt *et al.*, 2012) to take these differences into account (see Supplementary Methods). This approach assigns a "domain score" that can be used to rank or filter inferred SLiM-domain interactions. Once this step is completed, the inferred interactions are stored in both tab-delimited and JSON files to facilitate the import in other applications, such as Cytoscape (Shannon *et al.*, 2003).

In the final step, in order to identify the host cellular functions potentially targeted by the pathogen proteins, *mimic*INT executes a functional enrichment analysis of host inferred interactors. This analysis statistically assesses the over-representation of functional categories, such as Gene Ontology terms and biological pathways (e.g., KEGG and Reactome), using the g:Profiler R client (Raudvere *et al.*, 2019).

Given the degenerate nature of SLiMs (Davey *et al.*, 2012), their detection is prone to generate false positive occurrences. For this reason, we implemented an optional sub-workflow that, using Monte-Carlo simulations, assesses the probability of a given SLiM to occur by chance in query sequences and, thus, can be used to filter out potential false positives (Hagai *et al.*, 2014) (see Supplementary Methods).

To ease deployment and ensure reproducibility and scalability on high-performance computing infrastructures, *mimic*INT is provided as a containerized application based

on Docker and Singularity (Merkel, 2014; Kurtzer *et al.*, 2017). *mimic*INT is available at https://github.com/TAGC-NetworkBiology/mimicINT.

Results

We sought to evaluate the ability of *mimic*INT to correctly infer SLiM-domain interactions, as this inference can generate many false positives (Weatheritt *et al.*, 2012), using the default parameters for SLiM detection (see Implementation). To do so, we used as controls two datasets of established motif-mediated interactions (MDI) from the ELM database (Kumar *et al.*, 2020) (see Supplementary Methods): *(i)* 103 interactions between 87 viral and 44 human proteins (vMDI); *(ii)* 31 interactions between 16 bacterial and 23 human proteins (bMDI). We were able to correctly infer the majority of these interactions (91 vMDI, true positive rate = 88.3%; 21 bMDI, true positive rate = 67.7%). As the availability of negative SLiM-mediated interaction datasets is very limited (Weatheritt *et al.*, 2012; Idrees *et al.*, 2018; Kumar *et al.*, 2020), we estimated the false positive rate (FPR) by applying *mimic*INT to two sets of randomly generated interactions sets (degree-controlled, $vMDI_{rnd}$ and $bMDI_{rnd}$, respectively). Thirty-four $vMDI_{rnd}$ and 7 $bMDI_{rnd}$ were inferred as motif-mediated (FPR = 33% and FPR = 23%, respectively). We next annotated the human proteins in the two random sets with domain similarity scores. We kept only interactions for which the domain score was above 0.4 (Weatheritt *et al.*, 2012), thereby reducing the number of random interactions predicted as motif-mediated to 9 (FPR = 8.7%) for $vMDI_{rnd}$ and 2 (FPR = 6.4%) for $bMDI_{rnd}$. Finally, we tested *mimic*INT on two sets of experimentally verified negative 37 viral-human and 4 bacterial-human protein

interactions from the Negatome 2.0 database (Blohm *et al.*, 2014). Only two virus-human interactions (5.4%) were inferred as motif-mediated by *mimic*INT.

In the light of these results, we used *mimic*INT to infer the interactions between human proteins and the Marburg virus (MARV), an emerging infectious agent for which experimental protein interaction data is scarce (23 interactions for VP24 protein in IMEx interaction databases (Orchard *et al.*, 2012)).

We downloaded MARV protein sequences (7 proteins, Proteome ID: UP000180448) from UniprotKB in FASTA format. For domain detection, we considered only InterProScan matches in MARV sequences and ran *mimic*INT with default parameters.

In total, we inferred 11,431 interactions between 7 MARV and 2757 human proteins (see Supplementary Data). The vast majority of the inferred interactions, namely 10,101, are motif-domain interactions (MDI, 7 MARV and 2324 human proteins), and the remaining 1,339 are domain-domain interactions (DDI, 5 MARV and 479 human proteins). Interestingly, we observed an significant enrichment of known targets of other viruses among inferred interactors (1096 human proteins, 39.7% of the total, odds ratio = 1.3, P-value = 1.8×10^{-8}, one-sided Fisher's Exact test) (Orchard *et al.*, 2012): 62 (13% of DDI interactors, odds ratio = 0.2, P-value = 1, one-sided Fisher's Exact test) are involved in 133 inferred DDIs, and 1059 (45% of MDI interactors, odds ratio = 1.3, P-value = 6.7×10^{-6}, one-sided Fisher's Exact test) participated in 4591 inferred MDIs. By setting a stringent cutoff of 0.4 on the domain similarity scores, the number of inferred MDI decreases to 2082 (7 MARV and 597 human proteins), while the proportion of known viral targets among human interactors slightly increases (i.e.,

50%, 299 proteins, odds ratio = 1.4, P-value = $4.7x10^{-5}$, one-sided Fisher's Exact test).

None of the 23 experimentally identified interactions of the MARV VP24 proteins were identified by *mimic*INT, probably due to the fact that they were detected by an affinity-based purification method (Pichlmair *et al.*, 2012), which is more suited to identify indirect protein associations rather than direct interactions (Snider *et al.*, 2015). However, 17 MARV inferred interactions (17 MDI and 4 DDI) are supported by experimental evidence in the closely related Zaire Ebola Virus (Orchard *et al.*, 2012; Batra *et al.*, 2018).

The functional enrichment analysis performed by *mimic*INT on the full list of inferred host interactors returned a list of 975 enriched annotations at FDR<0.01 (see Supplementary Data). We next filtered out the functional categories annotating less than 5 or more 500 proteins obtaining a list of 763 enriched annotations (241 GO biological processes, 63 GO Cellular components, 6 CORUM complexes, 130 KEGG and 237 Reactome pathways), which points towards cellular processes and pathways related to viral infection and immune system (see Supplementary Data), thus further reinforcing the biological relevance of the inferred interactions.

Conclusions

We present *mimic*INT, a computational workflow enabling large-scale interaction inference between microbe and host sequences. Given the increasing frequency of (re-)emerging infectious diseases, *mimic*INT can be instrumental to better understand

the molecular details underlying microbial infections and to identify proteins and interactions as candidate points for therapeutic intervention. Although we developed *mimic*INT as a tool to infer protein interactions at the microbe-human interface, the workflow can be used to infer interaction among human proteins as well, or applied to organisms whose proteins bear either domains or SLiMs participating in known interaction templates.

Acknowledgments

The authors thank Paul de Boissier for helping in the early development of the workflow. The authors are also grateful to the members of the DIME project for fruitful scientific discussions and advices. Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high performance computing resources.

Author contributions: Conceptualization: S.A.C, C.B., L.S. and A.Z. Methodology: S.A.C., L.S. and A.Z. Software: S.A.C., M.C., K.M., L.D., L.S. and A.Z. Formal Analysis: S.A.C. and A.Z. Investigation: S.A.C., M.B. and A.Z. Writing – original draft: S.A.C. and A.Z. Writing – review & editing: C.B., L.S. and A.Z. Visualization: S.A.C. and A.Z. Supervision: C.B., L.S. and A.Z. Project Administration: C.B., L.S. and A.Z. Funding Acquisition: C.B and A.Z.

Funding

This work was supported by: the JPI HDHL-INTIMIC action co-funded by the Agence Nationale de la Recherche [ANR-17-HDIM-0001, DIME]; France 2030, the French Government program managed by the French National Research Agency [ANR-16-CONV-

0001] and from Excellence Initiative of Aix-Marseille University - A*MIDEX; and the European Union's Horizon 2020 Research and Innovation Programme [Project ID 101003633, RiPCoN]. SAC received funding from the "Espoirs de la recherche" program managed by the French Fondation pour la Recherche Médicale (FDT202106013072).

References

Arnold,R. *et al.* (2012) Computational analysis of interactomes: current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space. *Methods*, **57**, 508–518.

Batra,J. *et al.* (2018) Protein Interaction Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication. *Cell*, **175**, 1917-1930.e13.

Becerra,A. *et al.* (2017) Prediction of virus-host protein-protein interactions mediated by short linear motifs. *BMC Bioinformatics*, **18**, 163.

Blohm,P. *et al.* (2014) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*, **42**, D396-400.

Blum,M. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*, **49**, D344–D354.

Davey,N.E. *et al.* (2012) Attributes of short linear motifs. *Molecular bioSystems*, **8**, 268–81.

Davey,N.E. *et al.* (2011) How viruses hijack cell regulation. *Trends in biochemical sciences*, **36**, 159–69.

Dosztányi,Z. (2018) Prediction of protein disorder based on IUPred. *Protein Sci*, **27**, 331–340.

Edwards,R.J. *et al.* (2020) Computational Prediction of Disordered Protein Motifs Using SLiMSuite. *Methods Mol Biol*, **2141**, 37–72.

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

Franzosa,E.A. and Xia,Y. (2011) Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10538–10543.

Garamszegi,S. *et al.* (2013) Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks. *PLoS Pathog.*, **9**, e1003778.

Gfeller,D. *et al.* (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol Syst Biol*, **7**, 484.

Guven-Maiorov,E. *et al.* (2020) HMI-PRED: A Web Server for Structural Prediction of Host-Microbe Interactions Based on Interface Mimicry. *J Mol Biol*, **432**, 3395–3403.

Guven-Maiorov,E. *et al.* (2017) Prediction of Host-Pathogen Interactions for Helicobacter pylori by Interface Mimicry and Implications to Gastric Cancer. *J Mol Biol*, **429**, 3925–3941.

Hagai,T. *et al.* (2014) Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. *Cell reports*, **7**, 1729–39.

Idrees,S. *et al.* (2018) SLiM-Enrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions. *PeerJ*, **6**, e5858.

Jones,P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Köster,J. and Rahmann,S. (2018) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **34**, 3600.

Kumar,M. *et al.* (2020) ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res*, **48**, D296–D306.

Kurtzer,G.M. *et al.* (2017) Singularity: Scientific containers for mobility of compute. *PLoS One*, **12**, e0177459.

Lasso,G. *et al.* (2019) A Structure-Informed Atlas of Human-Virus Interactions. *Cell*, **178**, 1526-1541.e16.

Merkel,D. (2014) Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, **2014**, 2.

Mosca,R. *et al.* (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res*, **42**, D374-379.

Nourani,E. *et al.* (2015) Computational approaches for prediction of pathogen-host protein-protein interactions. *Front Microbiol*, **6**, 94.

Orchard,S. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.

Paulsen,K. (2019) Optimising intrinsic disorder prediction for short linear motif discovery.

Pichlmair,A. *et al.* (2012) Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature*, **487**, 486–490.

Raudvere,U. *et al.* (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*, **47**, W191–W198.

Rose,P.W. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res*, **41**, D475-482.

Schleker,S. *et al.* (2012) Prediction and comparison of Salmonella-human and Salmonella-Arabidopsis interactomes. *Chemistry & biodiversity*, **9**, 991–1018.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Snider,J. *et al.* (2015) Fundamentals of protein interaction network mapping. *Mol Syst Biol*, **11**, 848.

Via,A. *et al.* (2015) How pathogens use linear motifs to perturb host cell networks. *Trends Biochem. Sci.*, **40**, 36–48.

Weatheritt,R.J. *et al.* (2012) The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics*, **28**, 976–982.

Yamauchi,Y. and Helenius,A. (2013) Virus entry at a glance. *J Cell Sci*, **126**, 1289–1295.

Zanzoni,A. *et al.* (2017) Perturbed human sub-networks by Fusobacterium nucleatum candidate virulence proteins. *Microbiome*, **5**, 89.

Figures

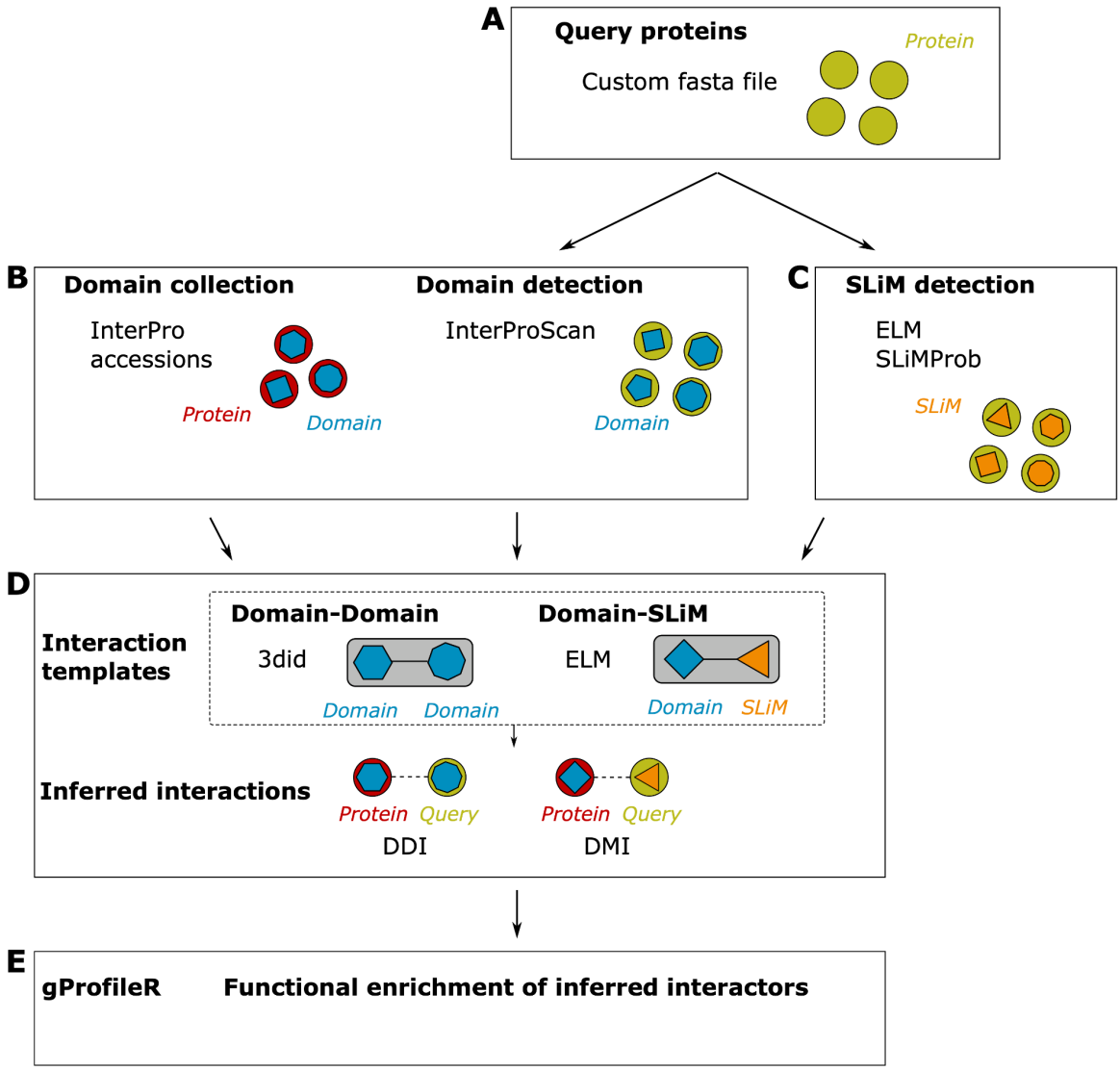

Figure 1: Overview of the *mimic*INT workflow. By providing a fasta file of protein sequences of the query species (e.g., microbe sequences) (A), *mimic*INT allows identifying both the domain (B) and SLiM (C) mediated interfaces of interactions. Using publicly available templates of interactions, *mimic*INT infers the interactions between the proteins of the query and target (i.e., host) species (D). Finally, it provides a list of functional annotations that are significantly enriched in inferred protein targets (E).