BACTERIA USE STRUCTURAL IMPERFECT MIMICRY TO HIJACK THE HOST INTERACTOME

Natalia Sanchez de Groot,[1,]* and Marc Torrent Burgas [2,]*

[1] Gene Function and Evolution Lab, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain, email: natalia.sanchez@crg.eu.

[2] Systems Biology of Infection Lab, Department of Biochemistry and Molecular Biology, Biosciences Faculty, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain, email: marc.torrent@uab.cat.

Keywords: protein interactions; bacteria; host recognition; protein structure

ABSTRACT

Bacteria use protein-protein interactions to infect their hosts and hijack fundamental pathways, which ensures their survival and proliferation. Hence, the infectious capacity of the pathogen is closely related to its ability to interact with host proteins. Here, we show that hubs in the host-pathogen interactome are isolated in the pathogen network by adapting the geometry of the interacting interfaces. An imperfect mimicry of the eukaryotic interfaces allows pathogen proteins to actively bind to the host's target while preventing deleterious effects on the pathogen interactome. Understanding how bacteria recognize eukaryotic proteins may pave the way for the rational design of new antibiotic molecules.

INTRODUCTION

In nature, bacteria do not exist in isolation but within communities of multiple species that require the bacteria to communicate and organize [1]. In particular, pathogenic bacteria need to interact with host cells to ensure their survival and proliferation [2]. Most of these interactions are mediated by protein effectors that are involved in microbial virulence [3]. These effectors, delivered to the host by secretion systems [4], outer membrane vesicles [5] or other specific mechanisms [6], allow bacteria to bind to host cells, replicate and spread within the host and subvert its immune system.

The collection of all these interactions between the host and the pathogen are known as the host-pathogen interactome. In the interactome, proteins with high connectivity degree are known as interactome "hubs" and are associated to essential and conditional phenotypes [7, 8]. The correlation between node degree and gene essentiality is known as the centrality-lethality rule [9] and has been observed in many organisms and interspecies interactomes [10]. In infectious diseases, pathogens selectively target hubs in the host interactome to rewire specific pathways for their own benefit [7]. The elucidation of these interactions is of outmost importance to understand how pathogens hijack the host systems.

Despite recent advances in the characterization of host-pathogen interactions, the structural knowledge of these protein complexes is very limited and largely restricts our ability to understand pathological systems. Here, we analyzed the degree centrality of bacterial proteins in the pathogen and the host-pathogen interactomes and investigate the structural characteristics of the interactions involved. We observed that hubs in the host-pathogen interactome are largely isolated in the pathogen interactome. This behavior can be explained by an imperfect mimicry of host interfaces that allow bacteria to minimize the toxicity of these proteins by restricting the number of interactions while maintaining affinity for host proteins.

RESULTS

Hubs in the host-pathogen interactome are segregated in the pathogen interactome

Because effectors target specific host interactions, their interfaces need to be fine-tuned to interact with a precise set of targets [11]. However, the number of protein folds [12] and interface geometries [13] is limited and these structures are frequently 'reused', allowing a single protein to bind many partners. If proteins are not controlled for promiscuous interactions, this can lead to toxic effects [14]. In the case of effectors, optimized interfaces must be controlled to avoid promiscuous pathogen self-interactions that may compromise cell performance.

A safe ward strategy against unwanted interactions is the timely expression of proteins under certain circumstances [15]. If bacterial effectors were only produced when bacteria are in close contact with the host, their deleterious effects could be minimized. However, this does not seem to be the case. The expression of effectors delivered by secretion systems is not triggered upon infection in cases where data is available (**Figure S1A, Supporting Information**). Also, other mechanisms of delivering virulence factors, such as extracellular vesicles, encapsulate large amounts of proteins from the bacteria cytoplasm and periplasm and do not specifically select their cargo [16]. This strategy would also be non-optimal as detection and killing of bacteria by the innate immune cells is very fast [17] while altering the expression of a protein may take longer [18].

In this context, either these proteins are physically isolated form the rest of the proteome or are integrated in the network in a controlled manner. In eukaryotes, spurious interactions are prevented by segregating proteins into different compartments. However, bacteria lack these compartments. A possible alternative would be the use of protein condensates, non-membrane bound structures formed by liquid-liquid phase separation [19]. These condensates would allow proteins to be relatively isolated from the cell milieu, but ready to be delivered when required. However, we did not find a clear difference in condensate propensity between secreted effectors and the rest of the proteome (**Figure S1B, Supporting Information**). Besides, bacteria are more proficient in forming solid aggregates, that are less dynamic than liquid condensates [20].

Based on these evidences, we hypothesized that effectors could be integrated in the bacteria network in such a way that they were not deleterious when pathogens replicate outside the host but can be effectively deployed upon infection. By isolating effectors from the pathogen network, they would have less control over the interactome, minimizing the side effects. In agreement with this reasoning, we observed that effector proteins are significantly depleted of pathogen hubs (**Figure 1A**).

To further investigate the integration of effectors in protein networks we compared its degree centrality both in the pathogen and the host-pathogen interactomes (**Figure 1B**). We were able to separate proteins into three clusters: (*i*) proteins that have a high number of interactions in the host-pathogen interactome but are highly isolated in the pathogen interactome (C1 cluster, **Figure 1C**), (*ii*) proteins that are isolated in the host-pathogen interactome but largely connected in the pathogen interactome (C2 cluster, **Figure 1C**) and (*iii*) proteins that are mainly isolated in both interactomes (C3 cluster, **Figure 1C**).

We investigated whether the structural properties of these proteins may explain the differences observed between clusters (**Figure 1D**). We found that the proteins belonging to the C1 cluster are richer in coil structure compared to proteins in clusters C2 and C3. This is consistent with an increased propensity to disordered regions, a property commonly observed in eukaryotic proteins. The proteins that define the C2 cluster tend to form alpha helix, which favors the binding to nucleic acids. Finally, the proteins belonging to the C3 cluster are enriched in beta sheet structure and more prone to aggregate. Overall, these clusters show specific structural properties and may reflect the differences in the cellular milieu between prokaryotes and eukaryotes.

The fact that we did not find any cluster with a high number of interactions in both interactomes suggests that protein interfaces in the host and the pathogen are nearly orthogonal, meaning that is difficult to optimize an interface to act as a hub in the pathogen and the host-pathogen interactomes at the same time. These observations raise a fundamental question: how can bacteria discriminate between self and non-self interfaces?

Bacteria use structural imperfect mimicry to interact with the host

Protein interactions are mediated by non-covalent bonds between residues located in the interaction core, which is a central area excluded from the solvent. This core region is surrounded by the rim area, which is partially buried, helps to exclude water molecules from the core and is involved in the modulation of the interaction. The core explains most of the binding energy of the complex, while the rim can tune the binding strength [21], particularly in small complexes [22].

To understand how interface structure can be used to discriminate self and non-self interfaces in bacteria, we mined the PDB for bacteria-eukaryote (BE) complexes and found 90 nonredundant entries (**Figure 2A**, **Table S1, Supporting Information**). We also selected 185 bacteria-bacteria (BB) and 687 eukaryote-eukaryote (EE) complexes for comparison (**Table S1, Supporting Information**). We divided proteins into three regions (interface, rim and surface) and analyzed the amino acid composition for each region. Overall, we could clearly distinguish these three regions based on their composition, with polar residues favored at the surface and hydrophobic residues more abundant in rim and interface regions (**Figure 2B**). Differences in each of these regions between BE, EE and BB complexes were more subtle (**Figure 2B-C**). While some residues (Trp, Phe and Lys) were enriched in the rim area in BE complexes, only differences in Leu composition were detected in the interface area (**Figure S2 and S3, Supporting Information**). This fact may be related to the higher conservation of the interface compared to the rim or surface regions [23]. It is also important to note that "affinity-defining" positions, located in the interface, are highly optimized whereas "specificity-defining" positions are usually non-optimal and are located at the rim area [11]. Hence, bacteria proteins may preferentially modify the rim area to discriminate between self and non-self interactions.

Despite maintaining the same composition at the interface level, the interaction pattern between amino acids was substantially different between complex types. We evaluated amino-acid interactions and compared the connectivity network of BE complexes with that of EE and BB complexes. In general, we found that the contribution, in terms of the number of interactions, for each amino acid in BE complexes was significantly correlated to EE but not to BB complexes (**Figure 2D**). The correlation between the network of interactions for each amino acid (**Figure 2E-F**) and their organization (**Figure S4, Supporting Information**) also confirmed that BE complexes were more similar to EE than BB complexes. These results support the theory that bacteria may use molecular mimicry to interact with host proteins. According to the mimicry hypothesis, bacteria can partly or completely imitate the structure of host proteins by mechanisms of gene transfer and/or convergent evolution using a strategy called 'molecular mimicry'[24]. Bacterial proteins competitively bind to the target host site [25] and redirect host hub proteins away from their pathway [26]. This strategy does not necessarily involve changing the entire structure of proteins but only certain residues in the interface or rim areas [26]. These bacterial proteins target host processes involved in cell adherence and invasion, which are essential for infection and explain why certain bacteria display strict host selectivity [27]. However, mimicry has been observed mostly on a case-by-case basis, using sequence or structure similarity [2, 28] or by solving isolated complexes [29].

While the evidence supporting structural mimicry is strong, we noticed clear differences between BE and EE complexes at the amino acid interaction level (**Figure 2E**). For example, Arg interactions had different preferences: Arg-Glu interactions were preferred in BE complexes, whereas Arg-Asp interactions were preferred in EE complexes. This might reflect an evolutionary adaptation, as Glu residues are preferred in eukaryotic interfaces compared to prokaryotic interfaces and vice versa for Asp residues (**Figure S2, Supporting Information,** p=0.016). In these lines, when analyzing directionality in BE interactions, we observed that some amino acids were frequently targeted at the bacterial interface (Tyr, Arg, Leu and Gly), whereas others were mainly targeted at the eukaryotic interface (Trp, Lys and Met; **Figure 2G**), being Trp was the most conspicuous case (**Figure S5, Supporting Information**). We noticed that Trp-Asp and Trp-Glu interactions were more common in BE complexes (25% of proteins had at least 1 anion-pi interaction) compared to the PDB interactome (less than 10% of proteins had at least 1 interaction) [30]. Asp and Glu were preferred in the bacterial interface, while Trp was mostly located in the eukaryotic interface, which coincides with Trp being more abundant in eukaryotes than bacteria (p-values 0.10 and 0.012, for core and rim, respectively). Furthermore, Trp in the eukaryotic interface had a higher contribution to complex stability compared to Trp in the prokaryotic interface (**Figure S6, Supporting Information**), suggesting that those interactions would contribute to complex stability. In almost all interactions in BE complexes (95%), Trp interacted with anionic residues through anion-pi interactions, which involves the contact of the negative density of Asp and Glu with the positive density at the edge of the aromatic ring (**Figure 2H**). Collectively, these results confirm that bacteria use molecular mimicry to interact with eukaryotic proteins, but also suggest that such mimicry is imperfect. Hence, although the composition of the central interface is similar across all complexes, the differences observed in its geometry can help discriminate between self and non-self interactions. Also, differences in the rim area would allow to further fine tune the binding. In the next

section, we explore the use of imperfect mimicry in the context of host-pathogen interactions.

Imperfect mimicry in the *Y. pestis – H. sapiens* interactome

During the course of infection, pathogens use proteins to rewire a myriad of biochemical processes that are required for efficient propagation [31]. We recently showed that pathogen proteins engaged in a higher number of interactions with the host also have a major impact on pathogen fitness during infection [7]. Hence, the relevance of pathogen proteins in the infection process is proportional to its ability to reorganize the host interactome. Unfortunately, complexes of bacterial proteins with human targets are largely underrepresented in the PDB database.

To further investigate this issue, we used the *Yersinia pestis-Homo sapiens* interactome and analyzed domain-domain associations (in terms of protein superfamilies) in comparison with the isolated *Y. pestis* and *H. sapiens* networks. Such an approach is justified because organisms mainly use the same 'building blocks' for protein interactions, and the function of domain pairs seem to be maintained during evolution [32]. We observed that an important number of associations are shared between the *Y. pestis-H. sapiens* interactome and the *H. sapiens* interactome (19%) compared with the *Y. pestis* interactome (0.72%, $p<0.00001$; **Figure 3A**). Consistently, the shared subnetwork (intersection) between BE and EE networks is more densely connected compared to the shared subnetwork between BE and BB networks (**Figure 3B-C**). Again, this suggests that the BE interactome is more closely related to the EE rather than the BB interactome.

To further validate these results, we filtered the *Y. pestis-H. sapiens* network with fitness data, which measures the relevance of a given bacterial gene in infection. Using this strategy, we created a subset of domain interactions that have a high impact on the fitness of *Y. pestis* during infection. The superfamily associations for such network are significantly enriched in domains related to infection (**Figure 3C, Table S2, Supporting Information**). When possible, we modeled the three-dimensional structure of the proteins involved in this network by sequence similarity and then obtained the structure of the complex by docking simulations (18 complexes). Docking procedures were not highly reliable to delineate interfaces in detail but helped to draw a coarse-grained view of the interactions. To investigate whether the predicted complexes are more similar to BB or EE complexes, the correlation coefficients for the interaction pattern of each residue were obtained (**Figure 3D**). Similar to previous results, we observed that the modeled interactions were more similar to EE complexes than BB complexes (**Figure 1F**). Overall, the correlation coefficients were lower when compared to those of the complexes deposited in the PDB, which can be attributed to the predicted nature of these complexes. Hence, although modeled data must be treated with caution, it reflects a general trend that is consistent with previous observations.

DISCUSSION

Based on the results presented here, we suggest that bacterial effectors have evolved their interfaces to imperfectly mimic eukaryotic complexes. During this process, effectors would have been subjected to two opposing forces: on the one hand, there would be an evolutionary pressure to increase the number of interactions with the host while, on the other hand, effectors would be forced to minimize the number of interactions with other pathogen proteins (**Figure 4A**). The first condition is necessary to increase the pathogen infectivity and its survival inside the host cells. The second one is less obvious but can be explained in terms of protein stickiness, which is higher in eukaryotes than prokaryotes. By mimicking eukaryotic interfaces, effectors become stickier, potentially increasing the number of spurious interactions with other pathogen proteins. Such poisonous interactions may compromise the cell viability; therefore, pathogens must find a balance between increasing infectivity and limiting toxicity. Using imperfect mimicry of eukaryotic interfaces, effectors are able to discriminate between bacteria-bacteria (self) and bacteria-host (non-self) interactions (**Figure 4B**).

The imperfect mimicry of protein interfaces has direct evolutionary consequences: as pathogen effectors mimic eukaryotic complexes to enhance adaptation, the host, in its turn, evolves its proteome to discriminate its own proteins from the pathogen mimics. This creates an arms race for survival between the host and the pathogen. Such behavior is observed in viral infections, particularly in those caused by poxviruses [33]. In a viral infection, host cells phosphorylate the eukaryotic initiation factor 2A (eIF2a) by protein kinase R (PKR) to inactivate translation. To restore translation, poxviruses evolved a protein called K3L that mimics eIF2a and competes with it for PKR phosphorylation. In its turn, primates also evolve eIF2a to discriminate it from K3L, creating a lasting evolutionary circle.

Effectors regulate pathogen adhesion, survival and proliferation in the host and so, they are frequently found to be essential for infection *in vivo* [34]. However, as mentioned before, these proteins are also isolated within the pathogen interactome, which explains why they are often classified as nonessential for the pathogen growth *in vitro* [7]. Unfortunately, most large-scale screening assays aimed to discover new antimicrobials are developed *in vitro*. Using such approaches, most effectors will never be discovered and the potential drugs that could be developed against them will remain unexplored. Our observations, therefore, confirm the need to redefine our discovery pipelines to properly reflect the host environment.

Last but not least, our results suggest that host-pathogen protein-protein interactions are potential targets for a new generation of antimicrobials. Because effectors use imperfect mimicry, molecules designed against them should target specifically the host-pathogen interfaces. Without disrupting the endogenous host interactions, these molecules should have limited side-effects. In summary, treatments interfering with the adhesion and invasion of bacteria to host cells could be used as preventive strategies during surgical procedures or after infection by reducing the resistance of pathogens to known antibiotics by combating their spread in the organism [35].

MATERIALS AND METHODS

Databases

The *Y. pestis* and *H. sapiens* interactomes were obtained from the String database[36]. Unless otherwise specified, only highly reliable interactions were included (combined score > 700). The *H. sapiens - Y. pestis, H. sapiens- B. anthracis* and *H. sapiens – F. tularensis* were downloaded from IntAct as reported in [37]. Specifically, the *H. sapiens-Y. pestis* interactome and contains 4,059 interactions from a random yeast-two-hybrid assay with a tenfold coverage of the coding capacity of *Y. pestis*. Fitness values were obtained from [38] using transposon sequencing (Tn-seq) and calculated as the ratio of the rates of population expansion for the two genotypes after infection of *Y. pestis* in a mouse model. In total, 1.5 million independent insertion mutants were screened with a coverage of ~70% of the *Y. pestis* genome. Protein superfamilies of *Y. pestis* and *H. sapiens* were obtained from UniProt. Structural parameters were obtained from [39] (alpha helix, beta sheet and coil propensity), [40] (aggregation propensity), [41] (disorder propensity) and [42] (nucleic acid binding propensity).

Interface definition and calculation of contact maps

The interface, rim and surface regions were calculated using a python script developed by the Oxford Protein Informatics group, which is freely available (http://www.stats.ox.ac.uk/~krawczyk/GetContacts.zip). Briefly, the interface residues were defined as those in close contact between two molecules in a given complex (4.5 Å). Rim residues were not engaged in intermolecular contacts but were close to the interface (contact between molecules < 10 Å) and can have a more subtle effect on binding. Surface residues were determined as residues not present in either the rim or interface region that display a surface accessible area greater than 20 $Å^2$. We considered that Trp residues were engaged in anion-pi interactions when the distance between the centroid of the aromatic ring and the anion was between 2-5 Å. Anion-pi distances and interaction angles (defined between 0° and 90°) were measured in Pymol. Contact maps were generated in R (version 3.4.4) using the function *cmap* included in the bio3d package [43].

Modeling and docking of *Y. pestis*-*H. sapiens* complexes

Proteins involved in complexes were retrieved from the PDB when possible. Otherwise, the three-dimensional structure was modeled using Modeller (version 9.21) [44] as long as a template had homology higher than 30%, spanning more than 75% of the protein length. Docking was performed using Frodock (version 2.0) with default parameters [45]. The interface for complexes ranked highest in the docking score was analyzed using the pipeline described before.

Network analysis

All protein networks were analyzed with Cytoscape (version 3.6.1) [46], and statistical calculations were performed in R (version 3.4.4). The degree (k) of a node i is defined as the number of edges linked to i. To compare the node degree between two networks,

we define the normalized degree as k/(n-1) where *n* is the number of nodes in the network. Betweenness centrality (C_b) was computed as follows:

$$C_b(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where *s* and *t* are nodes in the network different from *i*, σ_{st} denotes the number of shortest paths from *s* to *t*, and $\sigma_{st}(i)$ is the number of shortest paths from *s* to *t* that *i* lies on. The intersection of two networks was calculated using the merge function (intersection) in Cytoscape.

Statistical analyses

Unless otherwise specified, all p-values were calculated using the Mann–Whitney U-test and considered significant when $p<0.05$ (see Figure Legends for further details). The χ_2 test was used to determine whether there is a significant difference between the expected and the observed frequencies in two categories. In all cases, two-sided tests were used with a testing level $\alpha=0.05$.

REFERENCES AND NOTES

[1] C. Ratzke, J. Gore, Nat Microbiol 2016, 1, 16022.
[2] N. Drayman, Y. Glick, O. Ben-nun-shaul, H. Zer, A. Zlotnick, D. Gerber, O. Schueler-Furman, A. Oppenheim, Cell Host Microbe 2013, 14, 63.
[3] H. Yang, Y. Ke, J. Wang, Y. Tan, S. K. Myeni, D. Li, Q. Shi, Y. Yan, H. Chen, Z. Guo, Y. Yuan, X. Yang, R. Yang, Z. Du, Infect Immun 2011, 79, 4413; V. Memisevic, N. Zavaljevski, R. Pieper, S. V. Rajagopala, K. Kwon, K. Townsend, C. Yu, X. Yu, D. DeShazer, J. Reifman, A. Wallqvist, Mol Cell Proteomics 2013, 12, 3036.
[4] T. R. Costa, C. Felisberto-Rodrigues, A. Meir, M. S. Prevost, A. Redzej, M. Trokter, G. Waksman, Nat Rev Microbiol 2015, 13, 343.
[5] A. Guerrero-Mandujano, C. Hernandez-Cortez, J. A. Ibarra, G. Castro-Escarpulli, Traffic 2017, 18, 425.
[6] S. Guiral, T. J. Mitchell, B. Martin, J. P. Claverys, Proc Natl Acad Sci U S A 2005, 102, 8710.
[7] N. Crua Asensio, E. Munoz Giner, N. S. de Groot, M. Torrent Burgas, Nat Commun 2017, 8, 14092.
[8] H. Ahmed, T. C. Howton, Y. Sun, N. Weinberger, Y. Belkhadir, M. S. Mukhtar, Nat Commun 2018, 9, 2312.
[9] H. Jeong, S. P. Mason, A. L. Barabasi, Z. N. Oltvai, Nature 2001, 411, 41.
[10] K. Raman, N. Damaraju, G. K. Joshi, Syst Synth Biol 2014, 8, 73.
[11] M. Fromer, J. M. Shifman, PLoS Comput Biol 2009, 5, e1000627.
[12] C. Chothia, Nature 1992, 357, 543.
[13] M. Gao, J. Skolnick, Proc Natl Acad Sci U S A 2010, 107, 22517.
[14] T. Vavouri, J. I. Semple, R. Garcia-Verdugo, B. Lehner, Cell 2009, 138, 198.
[15] H. Ge, Z. Liu, G. M. Church, M. Vidal, Nat Genet 2001, 29, 482.

[16] M. J. Kuehn, N. C. Kesty, Genes Dev 2005, 19, 2645.
[17] W. A. Davies, J Reticuloendothel Soc 1983, 34, 131.
[18] G. W. Li, D. Burkhardt, C. Gross, J. S. Weissman, Cell 2014, 157, 624.
[19] E. A. Abbondanzieri, A. S. Meyer, Curr Genet 2019, 65, 691.
[20] M. Torrent, D. Pulido, M. V. Nogues, E. Boix, PLoS Pathog 2012, 8, e1003005; F. D. Schramm, K. Schroeder, K. Jonas, FEMS Microbiol Rev 2019.
[21] J. R. Brender, Y. Zhang, PLoS Comput Biol 2015, 11, e1004494.
[22] R. Agius, M. Torchala, I. H. Moal, J. Fernandez-Recio, P. A. Bates, PLoS Comput Biol 2013, 9, e1003216.
[23] M. Guharoy, P. Chakrabarti, Proc Natl Acad Sci U S A 2005, 102, 15447; E. Teppa, D. J. Zea, C. Marino-Buslje, Protein Sci 2017, 26, 2438.
[24] S. Sikora, A. Strongin, A. Godzik, Trends Microbiol 2005, 13, 522; N. C. Elde, H. S. Malik, Nat Rev Microbiol 2009, 7, 787.
[25] C. E. Stebbins, J. E. Galan, Nature 2001, 412, 701; P. Escoll, S. Mondino, M. Rolando, C. Buchrieser, Nat Rev Microbiol 2016, 14, 5.
[26] E. Guven-Maiorov, C. J. Tsai, R. Nussinov, Semin Cell Dev Biol 2016, 58, 136; A. Via, B. Uyar, C. Brun, A. Zanzoni, Trends Biochem Sci 2015, 40, 36.
[27] X. Pan, Y. Yang, J. R. Zhang, Emerg Microbes Infect 2014, 3, e23.
[28] A. C. Doxey, B. J. McConkey, Virulence 2013, 4, 453.
[29] D. Y. Lin, J. Diao, J. Chen, Proc Natl Acad Sci U S A 2012, 109, 1925.
[30] X. Lucas, A. Bauza, A. Frontera, D. Quinonero, Chem Sci 2016, 7, 1038.
[31] L. E. Reddick, N. M. Alto, Mol Cell 2014, 54, 321; A. P. Bhavsar, J. A. Guttman, B. B. Finlay, Nature 2007, 449, 827.
[32] Z. Itzhaki, E. Akiva, Y. Altuvia, H. Margalit, Genome Biol 2006, 7, R125; A. del Sol, P. Carbonell, PLoS Comput Biol 2007, 3, e239.
[33] N. C. Elde, S. J. Child, A. P. Geballe, H. S. Malik, Nature 2009, 457, 485.
[34] J. M. Rendon, B. Lang, G. G. Tartaglia, M. T. Burgas, Nucleic Acids Res 2019.
[35] S. M. Lehar, T. Pillow, M. Xu, L. Staben, K. K. Kajihara, R. Vandlen, L. DePalatis, H. Raab, W. L. Hazenbos, J. H. Morisaki, J. Kim, S. Park, M. Darwish, B. C. Lee, H. Hernandez, K. M. Loyet, P. Lupardus, R. Fong, D. Yan, C. Chalouni, E. Luis, Y. Khalfin, E. Plise, J. Cheong, J. P. Lyssikatos, M. Strandh, K. Koefoed, P. S. Andersen, J. A. Flygare, M. Wah Tan, E. J. Brown, S. Mariathasan, Nature 2015, 527, 323.
[36] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, C. von Mering, Nucleic Acids Res 2015, 43, D447.
[37] M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali, B. W. Sobral, PLoS One 2010, 5, e12089.
[38] S. G. Palace, M. K. Proulx, S. Lu, R. E. Baker, J. D. Goguen, MBio 2014, 5.
[39] G. Deleage, B. Roux, Protein Eng 1987, 1, 289.
[40] G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, M. Vendruscolo, J Mol Biol 2008, 380, 425.
[41] A. Campen, R. M. Williams, C. J. Brown, J. Meng, V. N. Uversky, A. K. Dunker, Protein Pept Lett 2008, 15, 956.
[42] A. Castello, B. Fischer, C. K. Frese, R. Horos, A. M. Alleaume, S. Foehr, T. Curk, J. Krijgsveld, M. W. Hentze, Mol Cell 2016, 63, 696.

[43] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, L. S. Caves, Bioinformatics 2006, 22, 2695.
[44] B. Webb, A. Sali, Methods Mol Biol 2017, 1654, 39.
[45] E. Ramirez-Aportela, J. R. Lopez-Blanco, P. Chacon, Bioinformatics 2016, 32, 2386.
[46] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Genome Res 2003, 13, 2498.

ACKNOWLEDGMENTS

Funding: MT would like to acknowledge support from the Programa Ramón y Cajal (RYC-2012-09999). This study was funded by the Ministerio de Economía y Competitividad (SAF2015-72518-EXP and SAF2017-82158-R) and a Research Grant 2016 by the European Society of Clinical Microbiology and Infectious Diseases (ESCMID), all to MT. NSdG acknowledges support from the Ministerio de Economía y Competitividad, 'Centro de Excelencia Severo Ochoa 2013-2017' and CERCA Programme from the Generalitat de Catalunya.

Author contributions: M.T.B. conceived and designed the experiments, M.T.B. and N.S.d.G conducted the experiments and wrote the manuscript. All authors read and approved the final text.

Competing interests: The authors declare no competing interests.

Data and materials availability: All data that support the findings reported in this manuscript are available online either in the links provided in the Methods section or supplied with this manuscript as Supplementary Information.

FIGURES AND TABLES

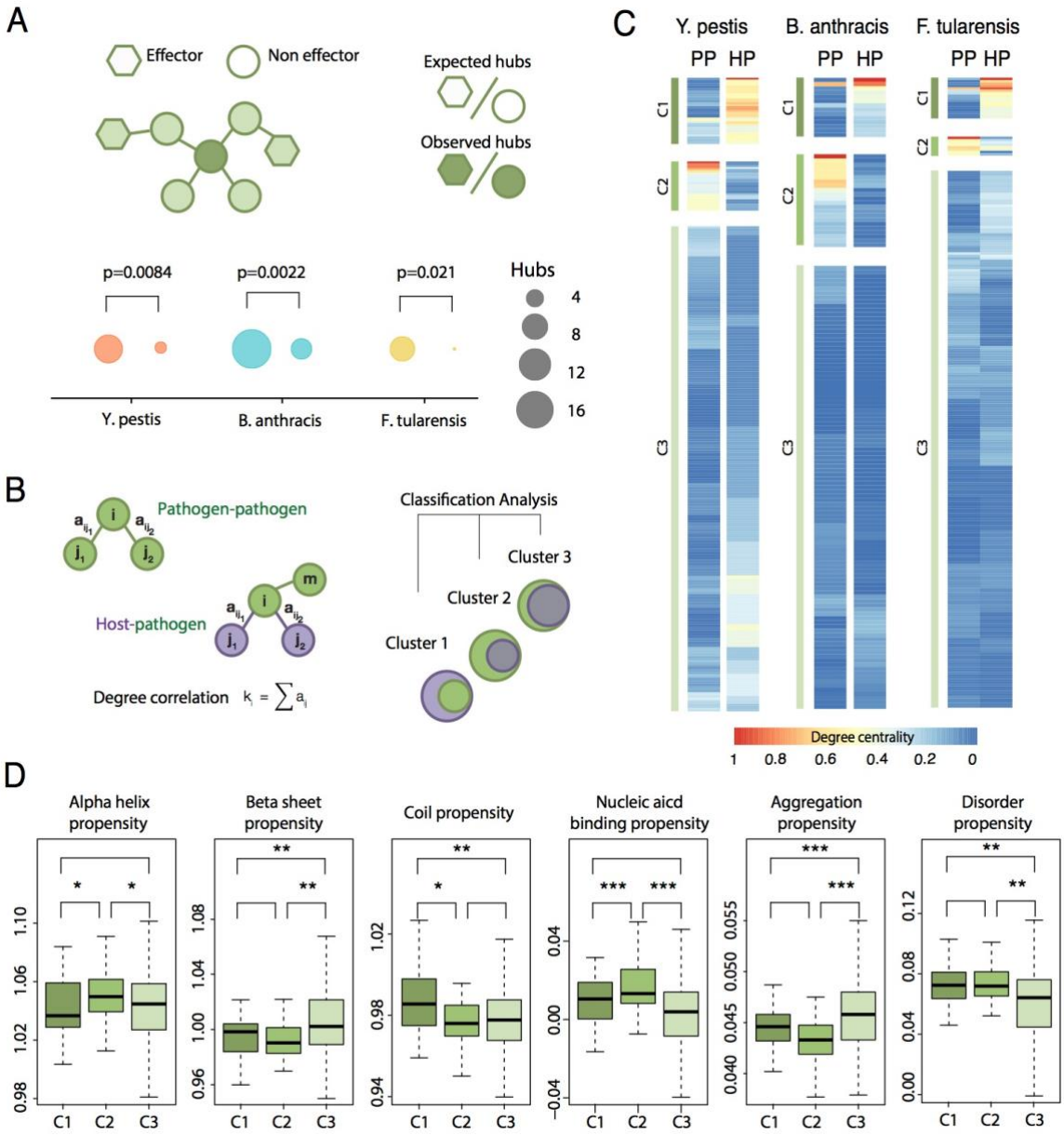

Figure 1. Hubs in the host-pathogen interactome are largely isolated in the pathogen network. (A) We computed the degree of interaction for all proteins in *Y. pestis*, *B. anthracis* and *F. tularensis* and asked whether the number of hubs (here defined as the 5% of most connected proteins) differ between effectors and non-effectors. In all cases, the number of hubs observed is significantly lower than expected. The p-values were computed using a χ^2-square test of independence to assess the probability of observing such a large discrepancy (or larger) between observed and expected values. (B) We compared the degree centrality of bacterial proteins in the pathogen and the host-pathogen interactomes. Based on the results obtained, we classified the bacterial proteins in clusters, according a k-means clustering algorithm. (C) Based on this clustering, three different groups were identified: proteins that have a high number of interactions in the host-pathogen interactome but are highly isolated in the pathogen interactome (C1 cluster); proteins that are isolated in the host-pathogen interactome but deeply connected in the pathogen interactome (C2 cluster) and proteins that are mainly isolated in both clusters (C3 cluster). (D) The three clusters

identified previously have distinctive structural properties. Proteins in C1 cluster are enriched in coil structure, which favors the presence of disordered regions; proteins in C2 cluster are enriched in alpha helix structure, which favors interaction with nucleic acids and proteins in C3 cluster are enriched in beta sheet structure, that favors aggregation. * $p<0.05$; ** $p<0.01$; ***$p<0.005$ using a Mann–Whitney U-test with $\alpha=0.05$.

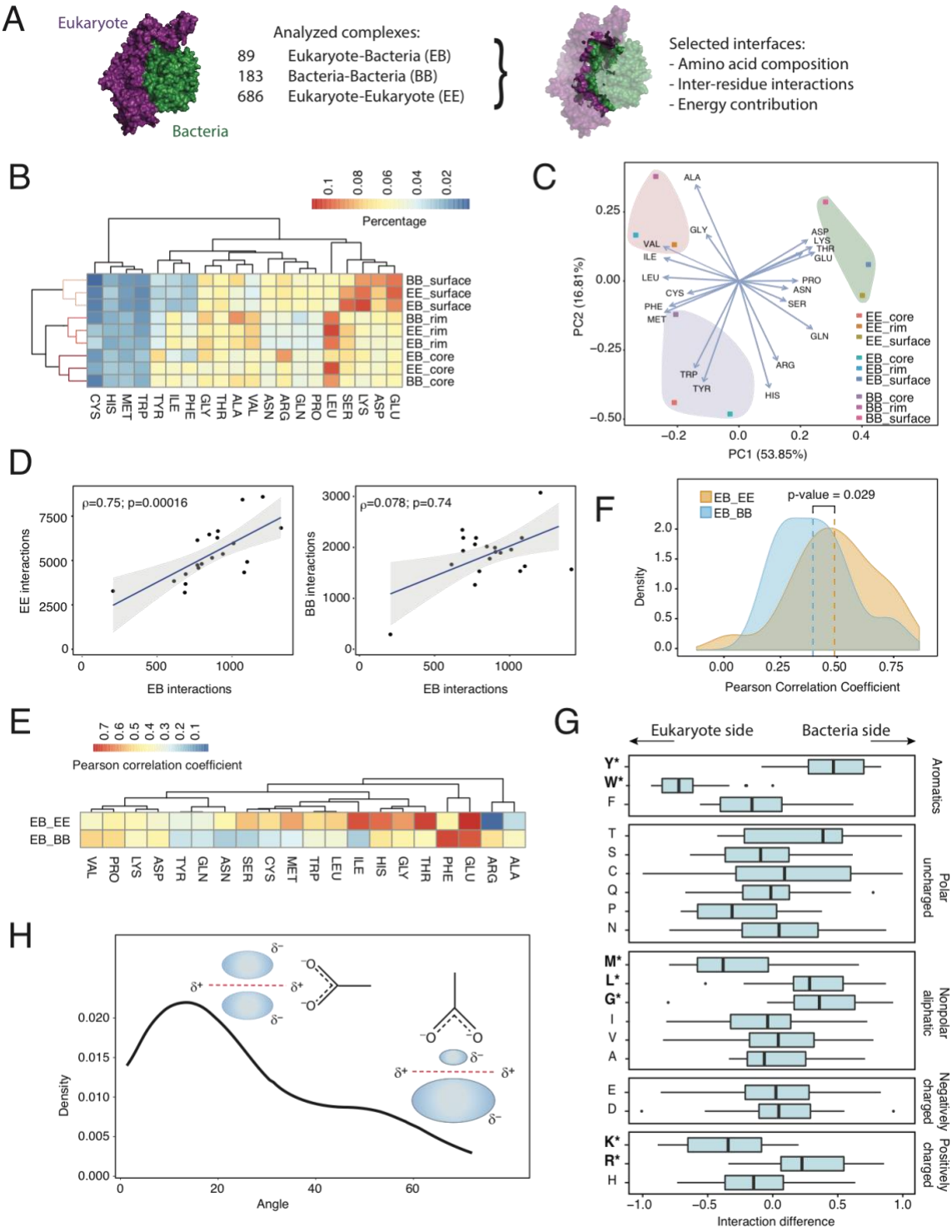

Figure 2. Analysis of protein complexes. (**A**), To compare the structural determinants of bacteria-eukaryote (BE) complexes, 90 nonredundant complexes were obtained from the PDB and compared to 185 bacteria-bacteria (BB) and 687 eukaryote-eukaryote (EE) complexes. (**B**) Hierarchical clustering and (**C**), principal component analysis of amino acid composition in BE, BB and EE protein complexes. (**D**) To characterize the interaction pattern in BE complexes, the number of interactions for each amino acid in BE complexes was plotted against EE and BB complexes. Regression lines were calculated using the Spearman rank-correlation approach to control for the impact of extreme values. (**E**) Hierarchical clustering of the interaction

pattern for each amino acid. The correlation coefficient for each amino acid was calculated comparing the number of interactions with all other amino acids in BE complexes against EE and BB complexes. (**F**) Distribution of Pearson correlation coefficients as calculated in panel F for BE complexes against EE and BB complexes. (**G**) Directionality for each amino acid (D_i) in BE interactions was calculated as the relative difference in the number of interactions (N) in both directions $D_i = \frac{N_i^B - N_i^E}{N_i^B + N_i^E}$., (**H**) Distribution of the angle measured for all anion-pi interactions involving Trp in BE complexes.

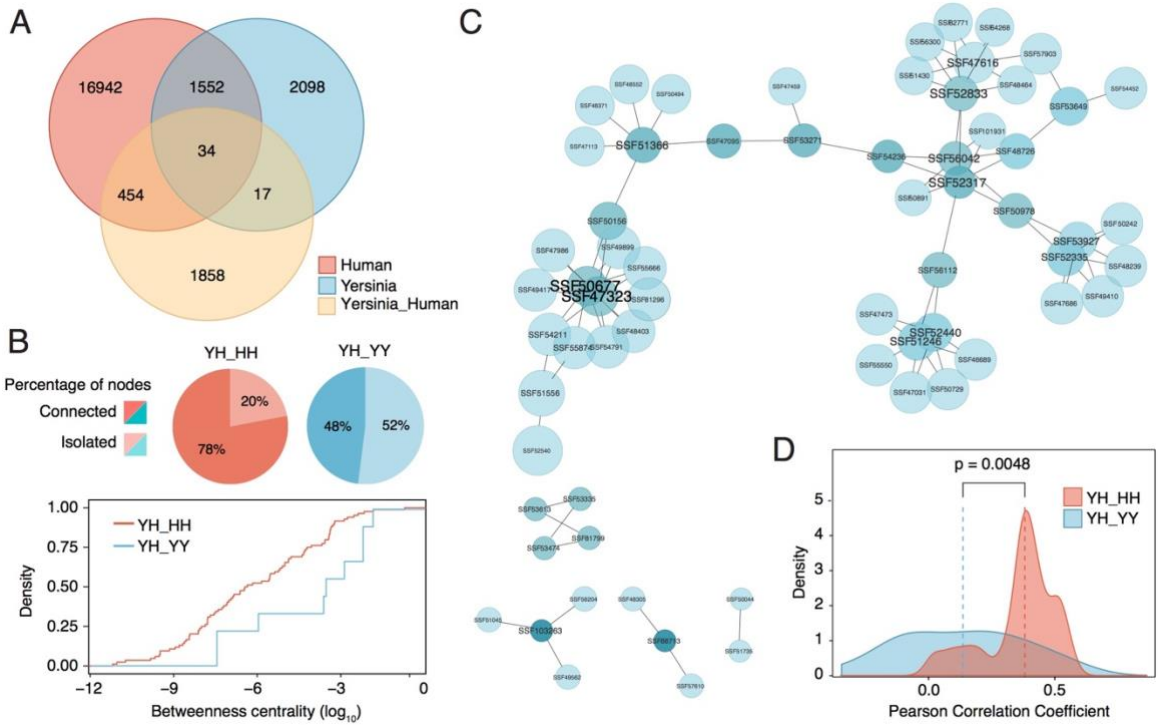

Figure 3. Structural analysis of protein-protein interactions in the *Yersinia pestis-Homo sapiens* interactome. (**A**) Venn diagram showing shared and unique domain associations between the *Y. pestis-H. sapiens*, *Y. pestis-Y. pestis* and *H. sapiens-H. sapiens* interactomes. (**B**) Percentage of isolated and connected nodes in the shared subnetworks (intersection) between the *Y. pestis-H. sapiens* interactome and the *Y. pestis-Y. pestis* or *H. sapiens-H. sapiens* interactomes. Cumulative distribution of betweenness centrality in both subnetworks. (**C**) *Y. pestis-H. sapiens* domain association network filtered for *Y. pestis* proteins that have a high contribution to infection fitness (fitness factor < 0.4). Complexes that involve bacterial proteins with a high contribution to infection fitness were modeled and docked to obtain the putative three-dimensional structure. (**D**) Distribution of correlation coefficients in the filtered network of contacts for all modeled complexes (n=18). The correlation coefficient for all amino acids was calculated comparing the number of interactions with each amino acid in BE complexes against EE and BB complexes.

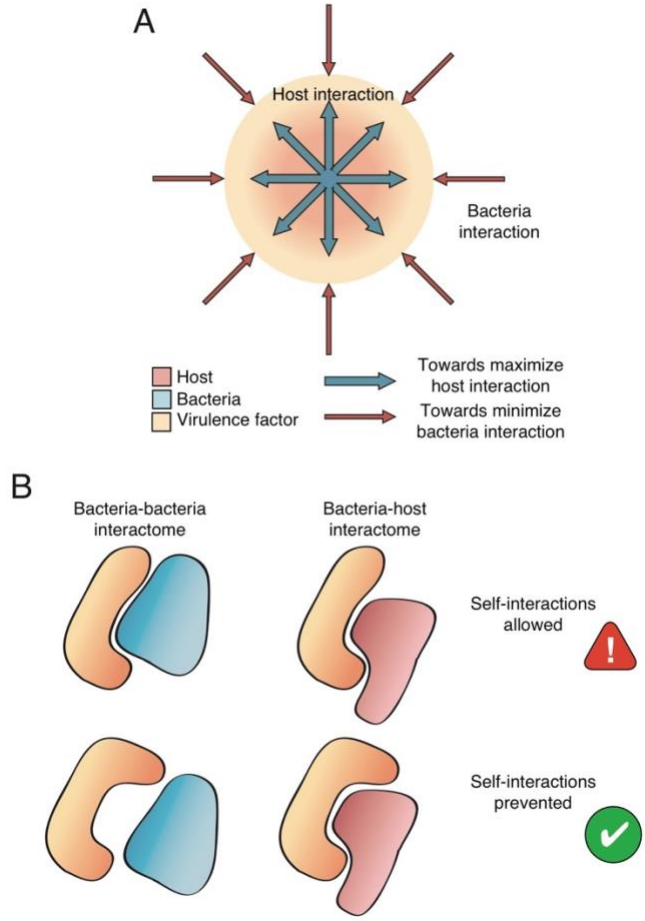

Figure 4. Organization restraints for host-pathogen interactions. (**A**) Effectors involved in host-pathogen interactions must optimize the interaction with host proteins while keeping undesired interactions under control within the pathogen interactome. (**B**) This balance is achieved through structural imperfect mimicry. Proteins retain the core interface to strongly interact with the host but modulate its geometry and the rim areas to minimize potentially detrimental self- interactions.

SUPPORTING INFORMATION

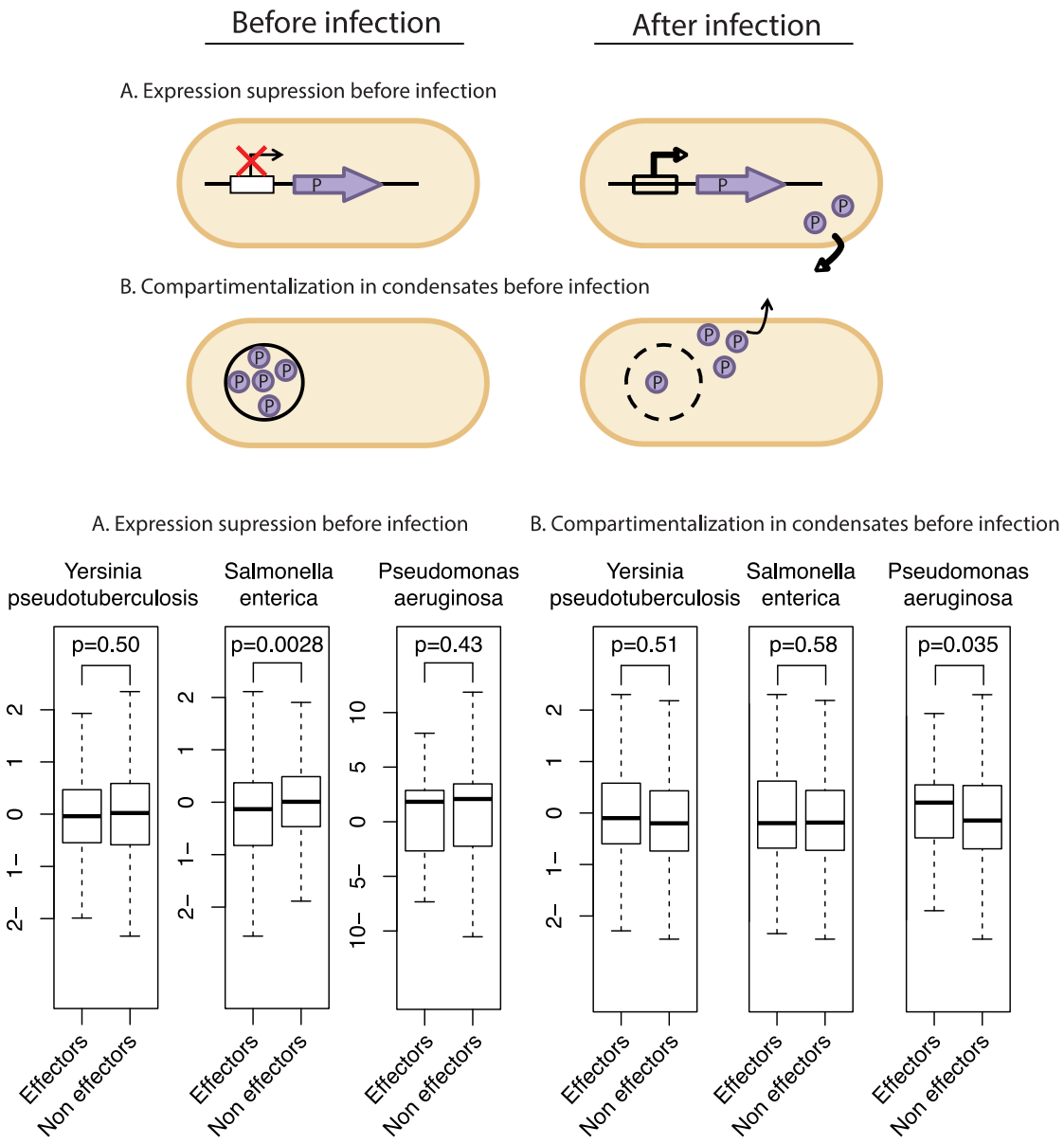

Figure S1. Strategies to control potentially detrimental interactions in the pathogen interactome. Pathogens could (A) regulate the expression of protein effectors or (B) compartmentalize the interactions in protein condensates for timely delivery upon infection. As observed in the boxplots in the lower panel, no clear significant differences were observed in all three organisms studied. P-values were calculated using the Mann-Whitney U test for comparing pairs of independent samples.

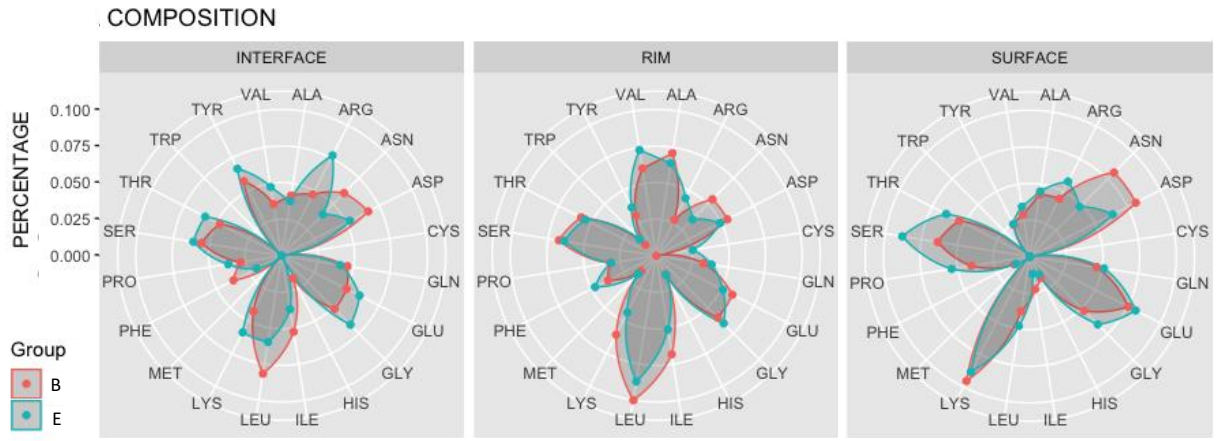

Figure S2. Amino acid composition of interface, rim and surface regions in bacteria-eukaryote complexes. Percentage of each amino acid in the interface, rim and surface regions of bacteria-eukaryote complexes. Each region is subdivided in bacteria (B, red) and eukaryote counterparts (E, blue).

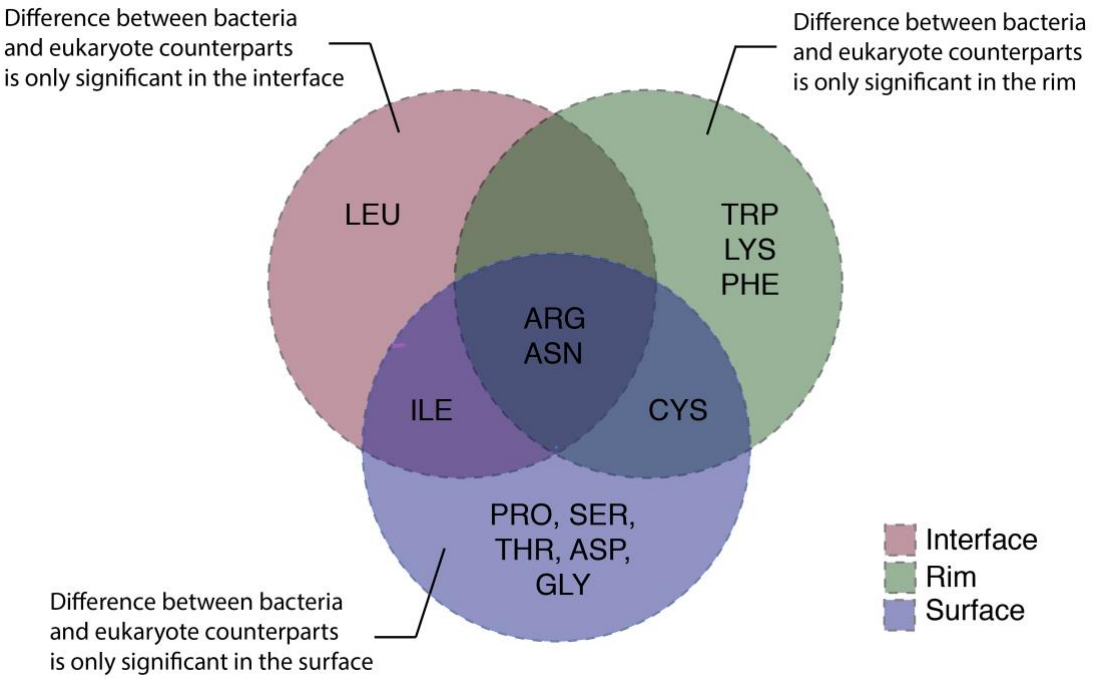

Figure S3. Amino acid preference in interface, rim and surface regions of bacteria-eukaryote complexes. Venn diagram displaying the differential amino acid composition of interface, rim and surface areas in bacteria-eukaryote complexes. Significant amino acid composition differences between bacteria and eukaryote counterparts in bacteria-eukaryote complexes were calculated using a Mann-Whitney U test and considered significant when $p < 0.05$.

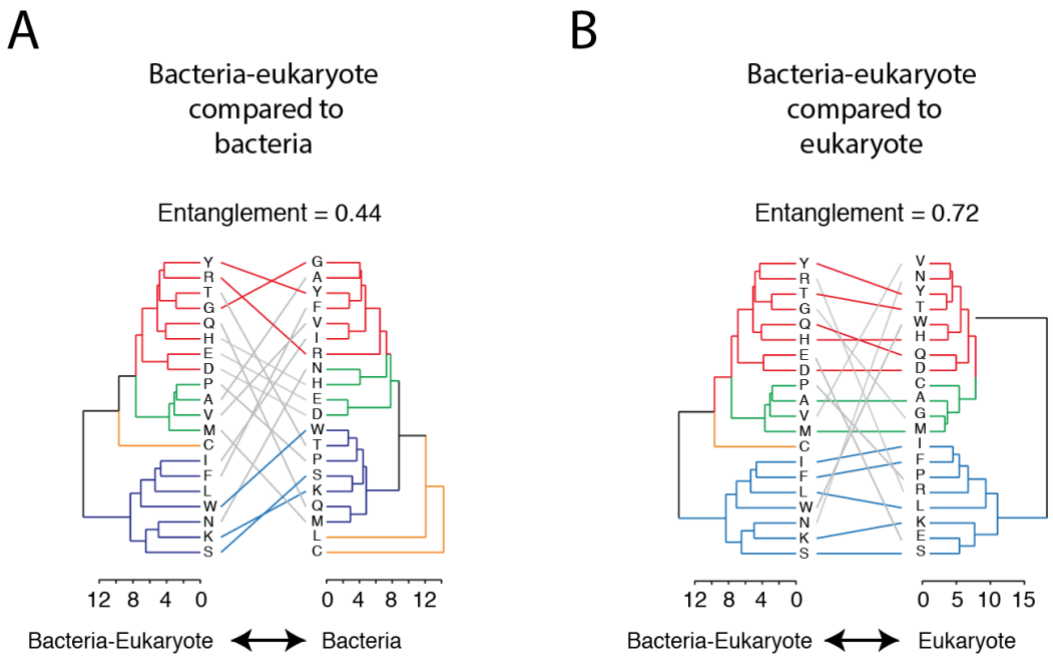

Figure S4. Comparison of interaction patterns in bacteria-eukaryote complexes compared to bacteria and eukaryote complexes. Dendrograms for each complex type (bacteria, eukaryote and bacteria-eukaryote) were built according to the interaction pattern of each amino acid using the Ward's minimum variance method. For each dendrogram, four groups (red, green, orange and blue) were defined using k-means clustering. Then, tanglegrams were built to compare the congruence between dendrograms (bacteria-eukaryote compared to bacteria and eukaryote). In the figure, congruence is depicted by the number of colored lines mapping common elements between same groups in different dendrograms. The quality of the alignment of the two dendrograms (entanglement) is also reported as a quantitative measure of congruence. Entanglement is measured between 1 (full entanglement, high congruence) to 0 (no entanglement, null congruence). Tanglegrams were built using the *dendextend* package in R1.

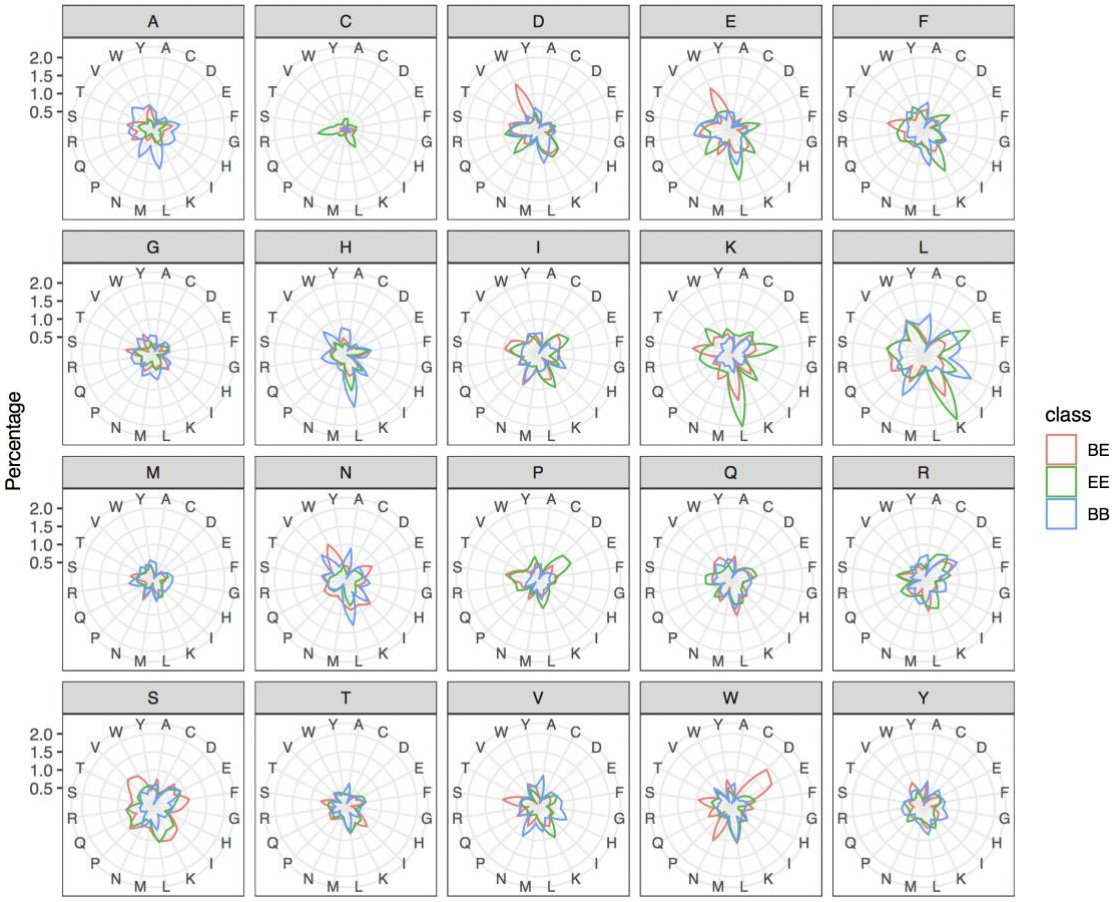

Figure S5. Amino acid connectivity analysis of bacteria-eukaryote, bacteria and eukaryote complexes. Number of interactions for each amino acid (measured as the percentage relative to the total number of interactions) in bacteria-eukaryote (BE, red), bacteria (BB, blue) and eukaryote (EE, green) complexes.

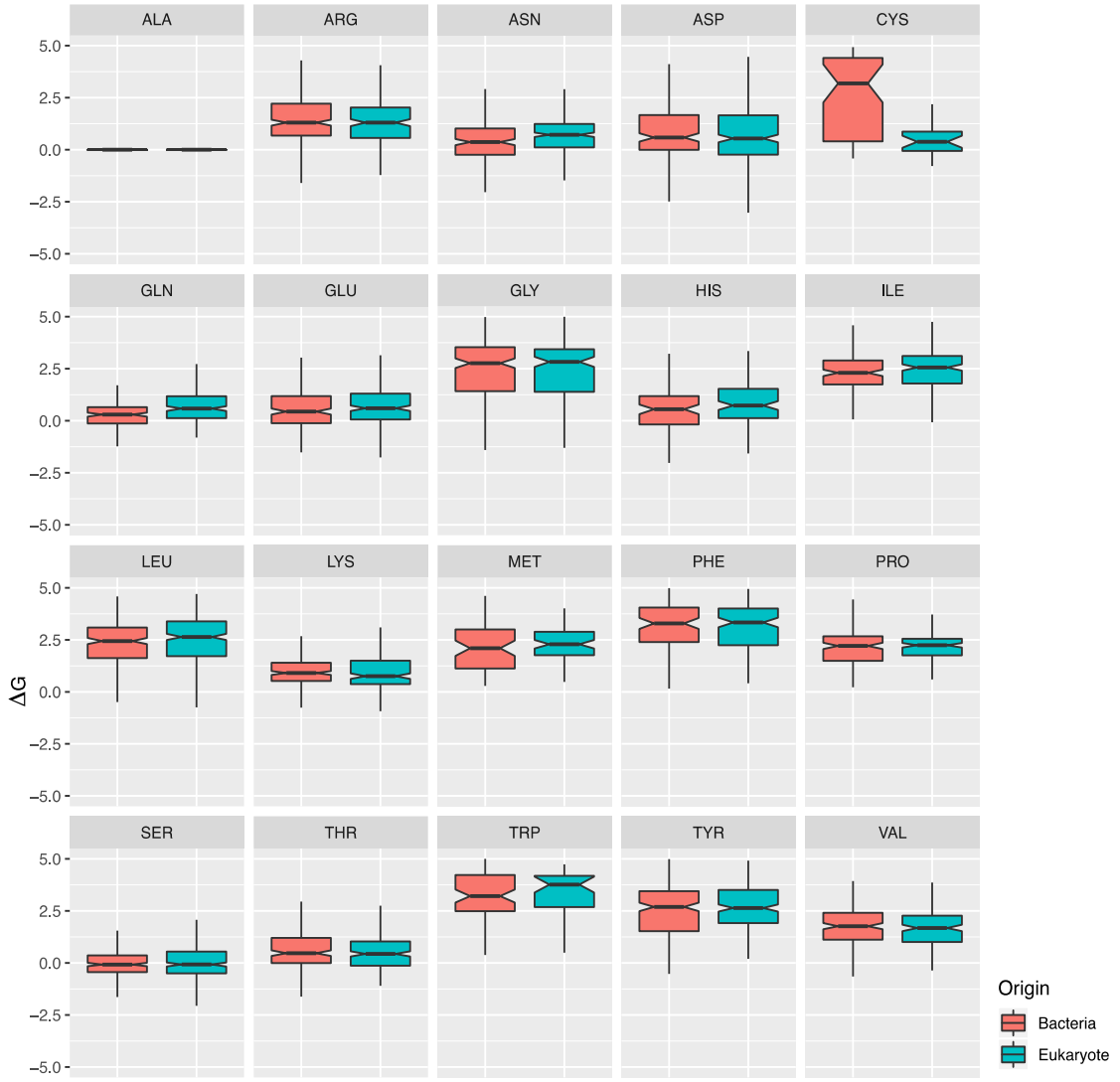

Figure S6. Residue contribution to the complex stability by alanine scanning in bacteria-eukaryote complexes. Each residue in the interface of bacteria-eukaryote complexes was mutated to alanine and the change in complex stability was calculated using FoldX2. The impact in complex stability for mutations in bacteria (red) and eukaryote (blue) interaction counterparts is compared.

Data file for Table S1. Uniprot codes of protein structures included in this study.

Data file for Table S2. List of superfamily associations significantly enriched in domains related to infection.

Supporting Information References

1 Galili, T. Bioinformatics 31, 3718-3720 (2015).

2 Schymkowitz, J. et al. Nucleic Acids Res 33, W382-388 (2005).