Computational and Structural Biotechnology Journal 20 (2022) 6534–6542

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

HPIPred: Host–pathogen interactome prediction with phenotypic scoring

Javier Macho Rendón, Rocio Rebollido-Ríos [1], Marc Torrent Burgas *

Systems Biology of Infection Lab, Department of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, Spain

ARTICLE INFO

Article history:
Received 15 August 2022
Received in revised form 9 November 2022
Accepted 10 November 2022
Available online 21 November 2022

Keywords:
Protein-protein interaction
Host
Pathogen
Bacteria
Pseudomonas aeruginosa
PAO1
interactome

ABSTRACT

Protein-protein interactions (PPIs) are involved in most cellular processes. Unfortunately, current knowledge of host-pathogen interactomes is still very limited. Experimental methods used to detect PPIs have several limitations, including increasing complexity and economic cost in large-scale screenings. Hence, computational methods are commonly used to support experimental data, although they generally suffer from high false-positive rates. To address this issue, we have created HPIPred, a host-pathogen PPI prediction tool based on numerical encoding of physicochemical properties. Unlike other available methods, HPIPred integrates phenotypic data to prioritize biologically meaningful results. We used HPIPred to screen the entire *Homo sapiens* and *Pseudomonas aeruginosa* PAO1 proteomes to generate a host-pathogen interactome with 763 interactions displaying a highly connected network topology. Our predictive model can be used to prioritize protein–protein interactions as potential targets for antibacterial drug development. Available at: https://github.com/SysBioUAB/hpi_predictor.

1. Introduction

During infection, pathogen proteins play a crucial role in rewiring multiple biochemical processes in the host, ultimately allowing the pathogen to attach to and invade host cells [1–3]. In return, the host uses sophisticated defense mechanisms against pathogens. Most of these processes are mediated by protein–protein interactions (PPIs) [4]. Several detection methods, such as yeast two-hybrid, pull-down assays, or coimmunoprecipitation, are commonly used to identify novel PPIs [5–7], although only a small fraction of the PPI space has been characterized so far [8]. In order to address the shortage of validated PPIs, in silico methods are commonly used.

There are many PPI predictors available, including homology-based [9,10], annotation-based [11,12], structure-based [13], and deep learning methods [14,15]. However, the potential of algorithms to predict host-pathogen PPIs is still far from optimal due to the lack of validated experimental datasets [8]. As a result, many of these algorithms display high false-positive rates [16], making them unusable for PPI discovery in the laboratory. To address these limitations, we have developed an algorithm to predict host-pathogen PPIs that uses numerical representations of proteins based on the physicochemical properties of their amino acids. To improve robustness, individual model predictions are then combined into a consensus interactome, which is finally integrated with phenotypic data collected from infection-related databases, allowing us to provide a ranked score for each interaction.

2. Methods

2.1. Data collection and dataset construction

Positive dataset: Host-pathogen PPIs were obtained from PHISTO [17], a database of experimentally validated interactions. A total of 9.237 inter-species PPIs between 95 different bacterial strains and *Homo sapiens* were used as a positive dataset. It is worth noting that 90 % of these entries belong to *Homo sapiens* – *Yersinia pestis* (4.069), *Homo sapiens* – *Bacillus anthracis* (3.053), and *Homo sapiens* – *Francisella tularensis* (1.348). Then, we applied a length cut-off to remove PPIs containing any protein shorter than

Abbreviations: PPI, Protein-protein interaction; CCC, Cross-correlation coefficient; BC, Betweenness centrality; FPR, False positive rate; ROC, Receiver-operating characteristic; PR, Precision-recall.

* Corresponding author.
E-mail address: marc.torrent@uab.cat (M. Torrent Burgas).
[1] Present address: University of Cologne, Faculty of Medicine, and Cologne University Hospital, Department of Internal Medicine, Center for Integrated Oncology Aachen, Bonn, Cologne, Düsseldorf, CECAD Cologne Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases, Cologne, Germany.

https://doi.org/10.1016/j.csbj.2022.11.026

100 amino acids or longer than 2.000 amino acids, obtaining a final dataset of 7.423 PPIs. Small proteins and peptides can easily correlate to small regions in other unrelated proteins, increasing the probability of false positive outcomes. By filtering entries shorter than 100 amino acids (~400 proteins for the pathogen proteome and ~700 for the host interactome) we avoid false positives while losing <0.5 % of potential interactions, on average. At the end, this dataset represents the interactome of 3.327 human proteins against 2.496 bacterial proteins.

Synthetic negative dataset: Protein libraries containing random sequences based on the 20-amino acid alphabet were created using a gamma distribution to fit the observed protein-length distribution in eukaryotic and bacterial organisms. The average protein lengths used were 472 for the eukaryotic proteome and 319 for the prokaryotic proteome [18]. The synthetic proteomes contained 20.000 and 3.000 proteins, similar in size to human and bacterial proteomes, respectively. Subsequently, the same length filtering criterion was used to remove proteins shorter than 100 or longer than 2.000 amino acids, giving a total set of 18.669 and 2.598 proteins for the host and bacteria, respectively.

Negative dataset: The *Homo sapiens* proteome was downloaded from UniProt (UP000005640) and used as host proteins. The combined proteomes of *Yersinia pestis* (UP000000815), *Francisella tularensis* (UP000001174), and *Bacillus anthracis* (UP000000594) were also downloaded from UniProt and used as pathogen proteins. Non-interacting pairs of proteins were generated by randomly pairing proteins from the host and the pathogen fractions, discarding pairs that belonged to the positive PPI dataset. The length filtering criterion was applied as previously described, to obtain a dataset of 7.421 entries, containing 2.734 and 2.102 different proteins from host and pathogen, respectively.

Query proteome datasets: *Homo sapiens* (UP000005640) and *Pseudomonas aeruginosa* PAO1 (UP000002438) proteomes were used to build the *Homo sapiens – Pseudomonas aeruginosa* interactome. After applying the length filtering criteria, host and bacteria proteomes were composed of 19.192 and 1.314 proteins, respectively.

2.2. Prediction of protein–protein interactions

The main steps involved in our prediction algorithm are: (1) the encoding of amino acid sequences to numerical strings, (2) the calculation of similarity scores between the query and positive datasets, (3) filtering entries by synthetic negatives, and (4) the prediction of putative PPIs based on their similarity scores (Fig. 1).

1. Numerical encoding of protein sequences: Each protein sequence in all datasets (positive, negative, synthetic, and query) was transformed into a numerical string by using physicochemical properties of amino acids, transforming the amino acid sequences into a numerical signal (Fig. 1A). The physicochemical properties are experimentally determined values for each amino acid as included in the AAindex database [19]. To represent the main properties that contribute to protein binding, we used five different physicochemical indices: alpha-helix propensity (GEIM800101) and beta-strand propensity (GEIM800105), to represent structure and ultimately hydrogen bonding; hydrophobicity index (ZIMJ680101) to represent hydrophobic effect; isoelectric point (ZIMJ680104) and electron–ion interaction potential values (COSI940101) to account for the electrostatic potential. All descriptors were normalized from 0 to 1 to avoid biases. After numerical encoding, a moving average algorithm with a sliding window of 9 positions was used to smooth the data and represent each amino acid's numerical value as a measure of itself and its near environment. Structural elements are, on average, between 5 and 10 (beta sheet) and 3–11 (alpha helix) residues long. Hence, using larger windows may result in information loss, while smaller windows would be less effective to capture the environment.

2. Calculation of similarity scores by cross-correlation: To determine similarity in the physicochemical profiles of proteins, we calculated the cross-correlation coefficients (CCCs) between the query dataset and the positive dataset by performing one-vs-all pairwise comparisons, i.e., each query protein was individually tested against all proteins in the positive dataset, for both host and pathogen proteins (Fig. 1B). To remove low-scored pairs, all comparisons with CCC < 0.4 were removed (Fig. 1C). Since CCCs depend on the lag parameter, we tested how many pairs were recovered by increasing the lag interval. For intervals higher than 200, no gain was observed (Fig. 2) so all CCCs were calculated with a lag interval of [-200, 200]. The highest CCC, i.e., {max (CCC), lag ∈ [-200,200]}, obtained for each pairwise comparison was assigned as a measure of similarity between the two proteins being compared, creating a database of similar proteins. Each entry in the database represents a pairwise comparison and includes information on the highest CCC and the length of the query protein.

3. Filtering the database using a synthetic negative dataset: To reduce the number of false positive predictions, we introduced a filtering criterion using the synthetic dataset. We evaluated the CCCs for all proteins in the synthetic set against host and pathogen proteins in the positive set. As the synthetic set contains only random-sequence proteins, none of them should be considered an interacting pair. The CCCs of the synthetic dataset were then plotted against their corresponding protein lengths. As the length of the proteins increased, their associated CCCs decreased linearly, a correlation that was also observed with the query datasets (Fig. 1C). Hence, we determined the slope and intercept points of parallel linear equations of the type $y = ax + b$ that represented such a negative linear correlation between protein length and CCC for the synthetic dataset, so that only 1 %, 0.1 %, 0.01 %, 0.001 %, 0.0001 or 0.00001 % of the data points fell above the equations. These linear equations were calculated and averaged over the five different physicochemical parameters previously described (Fig. 1C). Afterward, entries with CCC lower than or equal to the synthetic counterpart were discarded, to obtain a filtered database of similar proteins (Fig. 1C).

4. Prediction of protein–protein interactions: To predict new interactions, each PPI from the query dataset was inspected the following way: the host protein in the query PPI was searched against all the pairwise comparisons in the positive database of similar proteins and whenever a match was found, the query interaction was kept.

The same procedure was repeated for the pathogen protein in the subset of positive PPIs that contain the host match. If the query pathogen had a match in the subset, the interaction was considered a putative PPI (Fig. 1D). We performed this search sequentially with all the PPIs in the query dataset to obtain the predicted interactome (Fig. 1D). To generate consensus interactomes, individual host-pathogen interactomes were predicted for the five different physicochemical properties previously described and then combined to include any PPI that had been predicted by at least three individual models (Fig. 3A).

2.3. Phenotypic scoring of predicted PPIs

To add additional layers of information to the predicted interactions, we compared the proteins involved against several databases that contain information about the infection phenotype, namely BacFITbase [20], DualSeqDB [21], and PHI-base [22] (Fig. 3B).

1. Sequence alignment against BacFITbase: We downloaded BacFITBase v1.0 (accessed 2 September 2021), a database that contains information on bacterial fitness, as measured by transposon

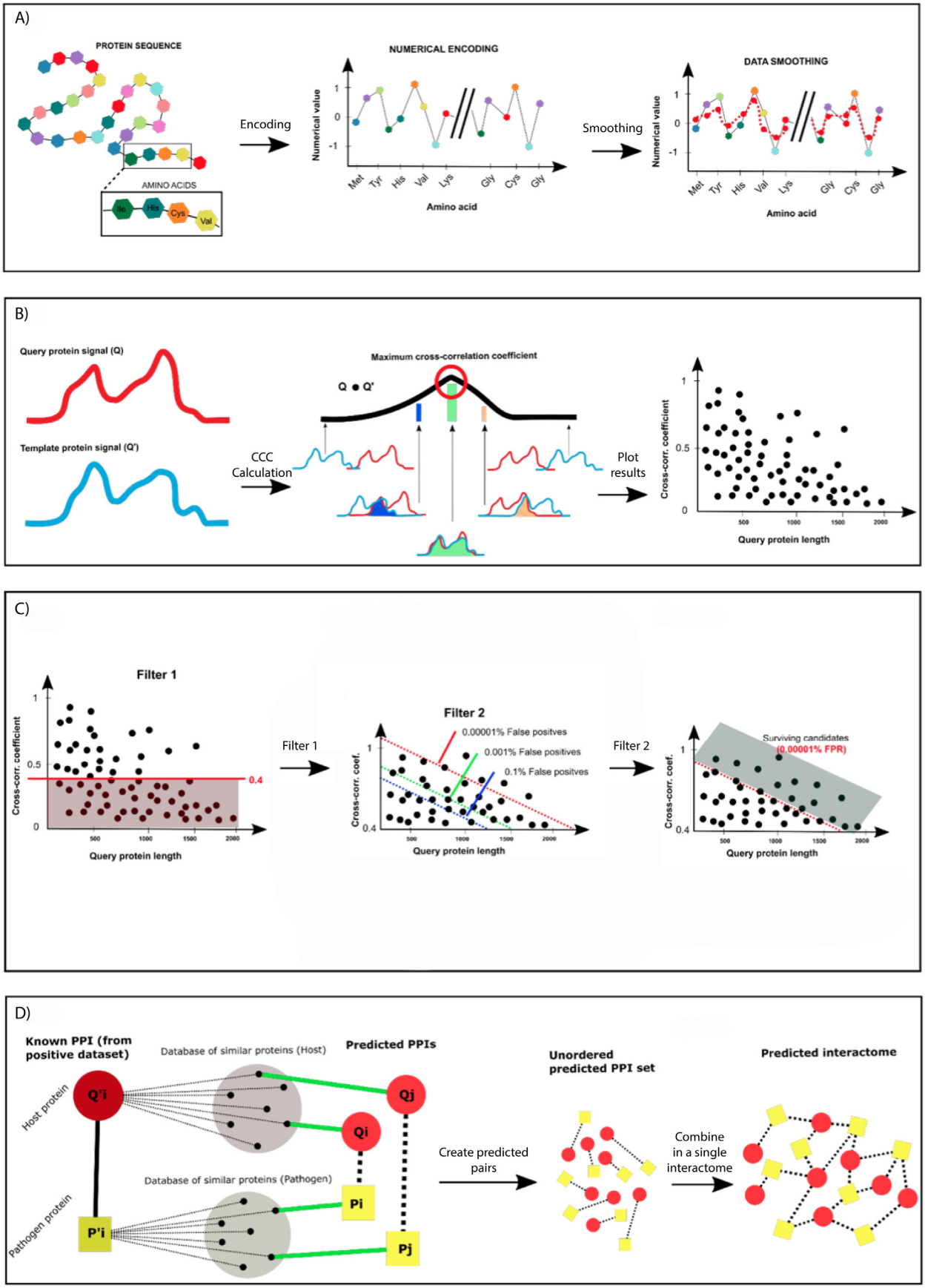

Fig. 1. PPI prediction algorithm using a single model. A) Protein sequences are numerically encoded using physicochemical descriptors and then smoothed using a sliding window approach. B) Protein profiles are then compared using cross-correlation and the calculated values are plotted against protein length. C) The dataset generated is first filtered by a coefficient cut-off (Filter 1) and later by a selected filter to control for false positive tolerance (Filter 2). D) At the end, the predicted protein pairs that survived the filtering steps are created and organized in a single interactome.

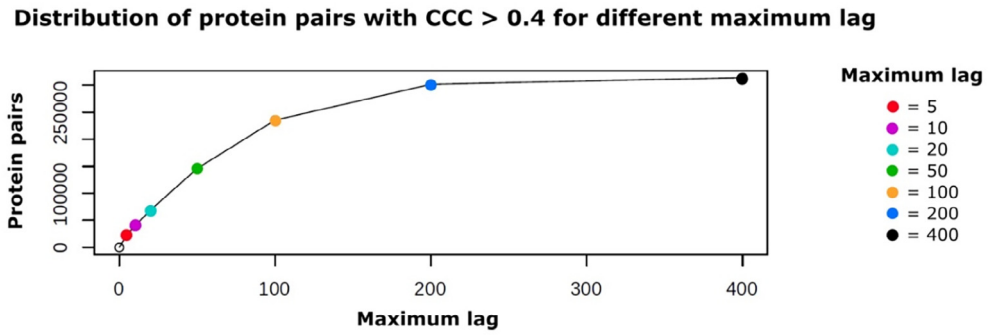

Fig. 2. Optimal maximum lag value for CCC calculation. To find the optimal maximum lag value, the CCCs between the *Homo sapiens* query proteins and the positive dataset were calculated for different lag intervals. The plot shows the number of protein pairs with a CCC > 0.4 for each maximum lag value tested for the hydrophobicity descriptor. The number of surviving protein pairs did not increase substantially after a maximum lag of 200.

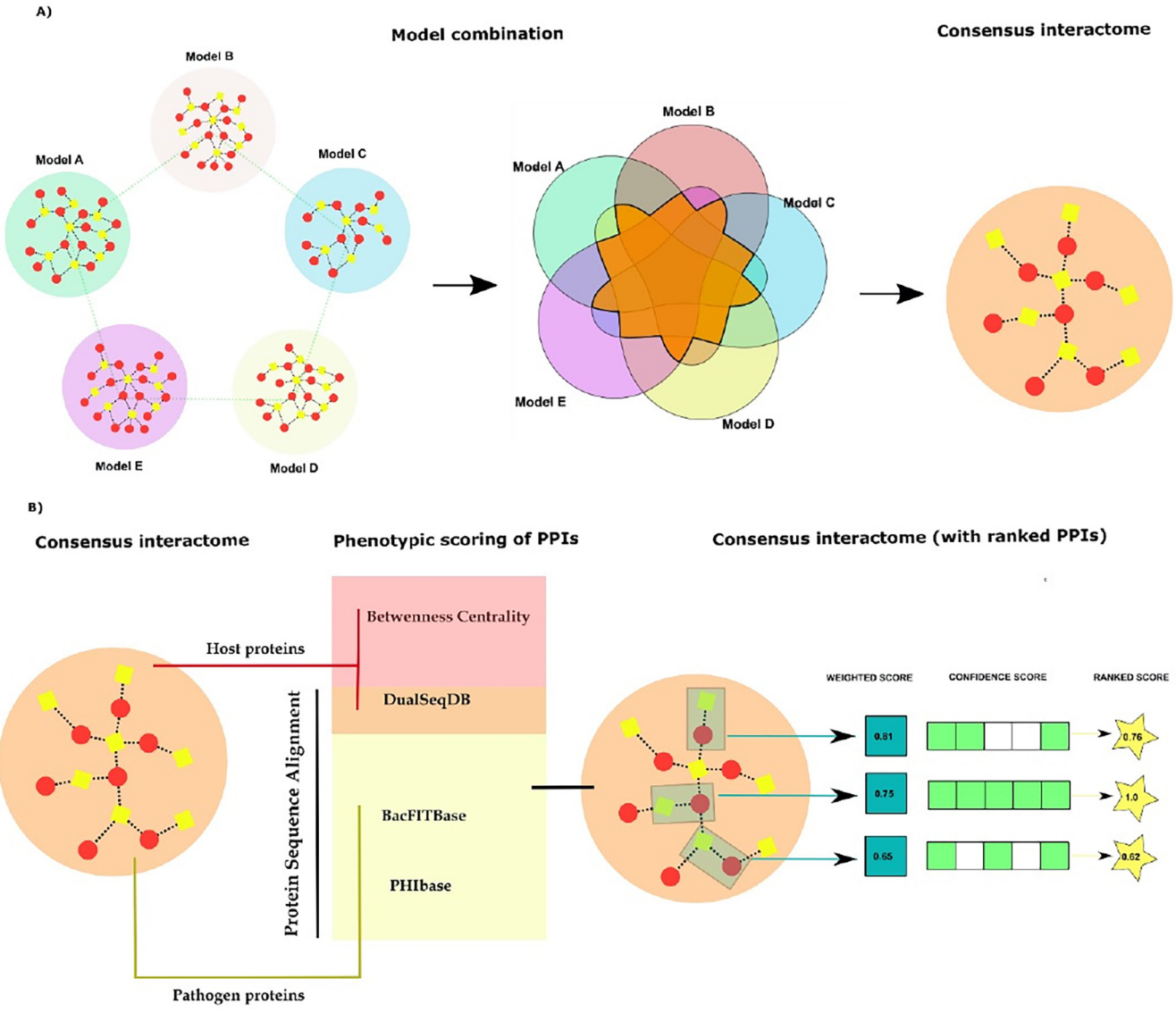

Fig. 3. Model combination and calculation of ranked scores. A) Model combination. B) Determination of ranked scores for the PPIs in the consensus interactome.

mutagenesis. To determine the similarity between pathogen proteins in the predicted interactome and entries in BacFITBase, HPIPred performs a BLAST sequence alignment between the query pathogen protein and the entire database, keeping all hits with a percentage of identity $\geq$ 40 % and an E-value < 10 that have a significant fitness score in BacFITBase (adjusted p-value $\leq$ 0.05). If multiple entries are retrieved, HPIPred assigns the average fitness score, and stores the mean standard deviation. Queries with no hits

are labeled as "NA". Finally, as BacFITbase assigns the lowest fitness scores to the most relevant proteins, the values are normalized from 0 to 1, assigning a value of 1 to the lowest fitness score reported, and a value of 0 to the highest.

2. Sequence alignment against DualSeqDB: We downloaded DualSeqDB 1.0 (accessed 2 September 2021), a database that contains information on gene expression changes in bacterial infection models. Changes in gene expression are represented as the $\log_2$ fold change, as measured by dual RNA-Seq experiments. HPIPred performs a protein sequence alignment between each query protein and DualSeqDB, for both the bacterial and host fractions, keeping those hits with a percentage of identity $\geq$ 40 % and an E-value < 10 that have a significant expression change score in DualSeqDB (adjusted p-value $\leq$ 0.05). The average $\log_2$ fold change is assigned to the query protein and the standard deviation is stored. Queries with no remaining hits are labeled as "NA". The standard 0–1 normalization is performed at the end, assigning a value of 1 to the highest reported fold change score, and a value of 0 to the lowest.

3. Sequence alignment against PHI-base: We downloaded PHI-base (accessed 2 September 2021), a dataset containing information on the role of pathogenic genes in bacteria. PHI-base assigns to each entry a "mutant phenotype", depending on how the gene deletion or mutation affects the pathogenicity of the organism. In some cases, the same gene can have more than one entry, since it may have been measured in different experiments. We filtered out those entries referring to pathogens that do not belong to the bacterial kingdom. Then, we only kept entries with mutant phenotype tags that matched "unaffected pathogenicity", "loss of pathogenicity", "reduced virulence", "lethal" or "increased virulence (hypervirulence)", and transformed them into numerical values, 0, 0.5 or 1 as follows: "lethal" = 1, "loss of pathogenicity" = 1, "reduced virulence" = 0.5, "increased virulence (hypervirulence)" = 0.5, "unaffected pathogenicity" = 0.

In the event that discrepant phenotypes were reported for the same database entry, the most abundant tag is assigned. HPIPred then performs a protein sequence alignment between each query protein and PHI-base. Hits with a percentage of identity $\geq$ 40 % and an E-value < 10 are retained. Surviving queries are assigned an average PHI-base score and the mean standard deviation is stored. "NA" labels are assigned to queries with no hits.

4. Betweenness centrality of host proteins: As suggested by the centrality-lethality rule [23,24], proteins that are central in the interactome are more likely to be essential for the organism. In this sense, betweenness centrality (BC) is a relevant centrality measure, as nodes with high betweenness are located on key communication routes and control network integrity [25]. Hence, we used BC as a proxy for protein relevance in the host. To measure protein essentiality in the *Homo sapiens* proteome, we calculated the BC score for all proteins. The *Homo sapiens* interactome was downloaded from the STRING database [26] (accessed 2 September 2021. We filtered out all PPIs with a confidence score lower than 0.9. We then used the R-package igraph [27] to build an undirected network, calculated the node BC score for all nodes in the graph, each representing a human protein, and performed a standard 0–1 normalization, being 0 the protein with the lowest BC score and 1 with the protein with the highest score (Fig. 3B).

5. Calculation of the ranked score: For each predicted PPI in the combined interactome we compiled all the normalized scores obtained in the previous steps (BacFITBase, DualSeqDB, PHI-base scores for the pathogen proteins, and betweenness centrality and DualSeqDB scores for the host proteins) and calculated an average score with a value ranging from 0 to 1 using the following equation:

$$AvS = \frac{F + E_h + E_p + P + BC}{NM} \quad (1)$$

where AvS is the average score, F is the fitness value, E_h and E_p are the $\log_2$ fold change in expression for host and pathogen, respectively, P is the infection phenotype, BC is the host betweenness centrality, and NM is the number of non-missing values. Then, a phenotypic weight (PW) weight was calculated to consider the number of missing values (NA) in the previous formula. Hence, a PW of 5 means no missing values, and 0 that no values were reported for that specific PPI. To account for both the average score and the confidence weight, we calculate a normalized ranked score (RS) (Fig. 3B):

$$RS = \frac{AvS * PW}{\max(AvS * PW)} \quad (2)$$

2.4. Model validation

To validate the performance of our algorithm, each individual PPI from the positive dataset was taken out of the predictive models and used as query input (leave-one-out cross-validation). All the PPIs recovered in the combined interactome generated were considered True Positives (TP), while the rest were considered False Negatives (FN). Subsequently, we passed each of the non-interacting pairs of proteins as input to our predictive algorithm. In this case, any non-interacting pair recovered in the combined interactome was treated as a False Positive (FP), whereas the remaining ones were treated as True Positives (TP). Evaluation metrics were calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (3a)$$

$$Recall = \frac{TP}{TP + FN} \quad (3b)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3c)$$

$$Sensitivity = \frac{FP}{FP + TN} \quad (3d)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3e)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (3f)$$

2.5. Software implementation

HPIPred was designed in R and wrapped up in a command-line tool that allows the user to choose the host and pathogen organisms. The software can be used in Linux and MacOS operating systems and is available at: https://github.com/SysBioUAB/hpi_predictor. Users can select a Uniprot proteome ID or upload the proteomes as custom files. They can also choose among more than 400 different physicochemical descriptors, select the preferred false positive rate and decide the number of models to build the consensus interactome. We also provide pre-calculated protein similarity values for five model organisms that can be used as hosts, namely *Homo sapiens*, *Mus musculus*, *Dario rerio*, *Caenorhabditis elegans, Drosophila melanogaster*, for the five default physicochemical descriptors. This option speeds up calculations when the model organisms are used. The calculation of an interactome

can take one day to compute in an average computer, while the pre-calculated proteomes can decrease the time to a few hours.

3. Results

3.1. Model prediction scores

One of the main goals in PPI prediction is keeping the number of false positives low. Hence, for algorithms built to predict inter-species interactomes, it is essential to control the false positive rate (FPR). Otherwise, the number of wrongly predicted interactions would bias the interpretation of the data and undermine the efforts to validate the results in the laboratory. When predicting bacteria-human interactomes, the number of potential interactions to test is around 80 million, assuming a bacteria proteome size of 4.000 proteins (20.000 human × 4.000 bacteria proteins). In this scenario, even with a small FPR of 0.1 %, the number of wrong predictions would be in the range of 80.000 interactions. In our method, we used synthetic proteomes, to keep these numbers low. These non-biological proteomes allowed us to make a coarse estimate of the FPR in large-scale predictions and control for low FPRs. Specifically, we set the maximum FPR to 0.0001 %, giving a theoretical estimate of 80 false positive interactions in a standard interactome.

To validate our predictions, we used the leave-one-out strategy. Hence, all entries in the positive dataset were used as templates to evaluate every interaction, except the one being tested. Using default settings, our method virtually achieves a precision of 100 % when combining all five models (Supplementary Information, Table S1). However, this comes at the cost of a low recall, i.e., only 2 % of the true interactions are recovered, as observed in the ROC and PR curves (Supplementary Information, Fig. S1). Although we achieved a low recall, the predictive power of our method is similar when compared with other protein–protein prediction algorithms [16], but with the advantage of controlling the FPR to very low levels (Fig. 4). In most sequence-based prediction algorithms, precision can drop to 10 % at 3 % recall and 20 % in annotation-based methods. All prediction methods, including our own, suffer mainly from using incomplete datasets of PPIs: since only a small fraction of the search space of PPIs has been experimentally validated, the models do not perform well for generalization, because the penalty of removing a known PPI from the positive dataset is very costly. Furthermore, for host-pathogen interactomes, only a few species of bacteria have been studied, which also limits the available information on which these predictors are based.

3.2. Prediction of the complete host-pathogen interactome Homo sapiens – Pseudomonas aeruginosa PAO1

To test how our model performs in a real case scenario, we predicted the *Homo sapiens* – *Pseudomonas aeruginosa* PAO1 interactome. The *H. sapiens* and *P. aeruginosa* proteomes contain 19.192 and 1.314 proteins, respectively, after length filtering, resulting in ~25 million putative interactions. We calculated the CCCs for all these interactions and generated the interactomes at 5 different FPRs (0.1 %, 0.01 %, 0.001 %, 0.0001 %, and 0.00001 %) with five physicochemical descriptors (GEIM800101, GEIM800105, ZIMJ680101, ZIMJ680104, COSI940101), as well as their combined model (Table 1). As noted before, in the case of single models, the allowed FPR can restrict the number of interactions from a hundred thousand PPIs when using the most permissive filter (0.1 %) to only several hundred PPIs with the most restrictive filter (0.00001 %). Also, the combination of different models can decrease the number of predicted PPIs but increase the confidence in the predicted interactions. The prediction of interactomes of relatively small sizes but with high confidence can be seen as an advantage for downstream analysis, including network visualization or gene ontology enrichment, but also for further experimental validation.

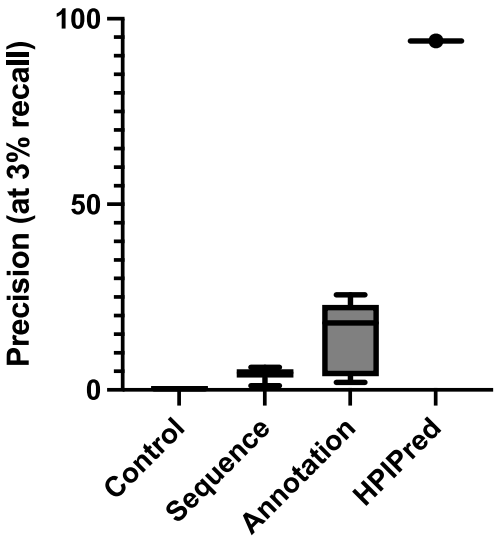

Fig. 4. Model evaluation of different PPI prediction methods. Sequence-based methods are based on numerical descriptors, annotation-based methods are based on domain and GO annotations, and HPIPred is the method described here. Control results were obtained based on equivalent sequence-based methods but using random numerical vectors instead of meaningful descriptors. Measurements for control, sequence- and annotation-based methods, were obtained from B. Dunham et al.[16].

We chose to further explore the predicted interactome generated with the most restrictive filter (FPR = 0.00001 %) due to its suitable proteome size (763 PPIs). After calculating the node betweenness centrality of the proteins in the *Homo sapiens* proteome and performing sequence alignment against DualSeqDB, BacFITBase, and PHI-base, we generated a ranked score for each PPI. This, in turn, allowed us to prioritize the interactions not only according to protein similarity based on physicochemical properties but also on network topology and biological properties related to the infection process. We visualized the predicted interactome in Cytoscape [28] (Fig. 4) by representing it as a bipartite graph, where host proteins (host nodes) are only connected to pathogen proteins (pathogen nodes) and vice versa. The ranked score of each predicted PPI was represented by the thickness and the color intensity of the edges, that is, the closer a ranked score to 1, the thicker its edge and the more intense its color, meaning that this PPI scored well on average on the biological databases.

In addition, we colored the nodes to represent changes in expression, derived from the BLAST search against DualSeqDB, in a range from blue (downregulated) to red (upregulated). We highlighted some of the top scoring PPIs in the Cytoscape network to show how highly ranked PPIs (Fig. 5), which represent protein pairs with inferred biological relevance, also appear to be important for network integrity and connectivity.

We then filtered out all PPIs with a ranked score lower or equal to 0.6 and divided the proteins involved in these PPIs into a pathogen fraction (16 proteins) and a host fraction (128 proteins) of unique proteins, which were then used to perform gene ontology enrichment analysis with DAVID [29] (Fig. 6). The results showed that the host fraction of proteins was enriched in biological processes related to immune and inflammatory response, such as regulation of NF-KappaB activity, cellular iron homeostasis, and actin filament bundle assembly. As for the pathogen fraction, we observed an enrichment in biological terms related to amino acid

Table 1
PPI sizes of the predicted interactomes by individual and combined models, at different FPRs.

Predicted interactions FPR	Model					
	ZIM101	ZIM104	GEIM101	GEIM105	COSI940101	Combined
0.1 %	142.089	386.529	551.043	143.493	225.456	2.594
0.01 %	8.469	15.155	22.907	7.082	15.119	1.661
0.001 %	1.853	2.322	3.491	1.485	3.717	1.104
0.0001 %	1.035	1.085	1.439	968	1.779	894
0.00001 %	814	915	915	763	1.146	763

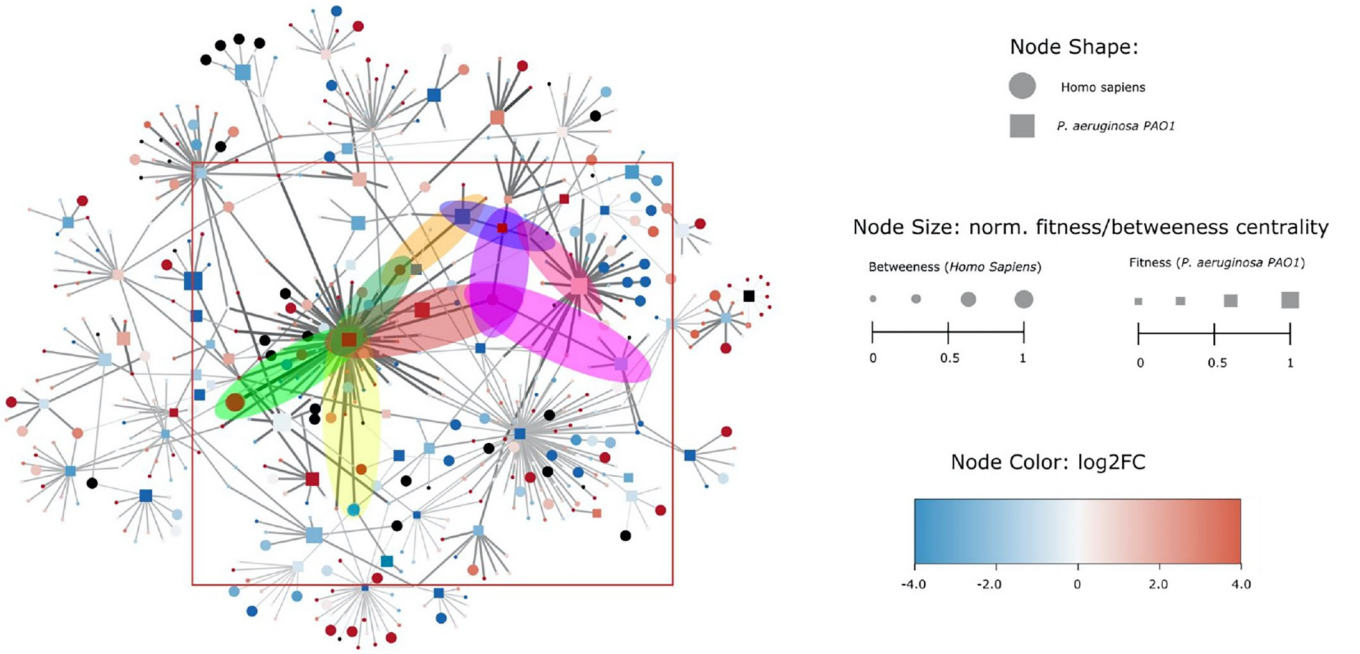

Fig. 5. Network representation of the *Homo sapiens – Pseudomonas aeruginosa* PAO1 interactome predicted by the combined models. Nodes represent proteins and edges represent predicted interactions between proteins. Host and pathogen proteins are represented by circles and squares, respectively. Nodes are colored according to the normalized expression changes (computationally derived from DualSeqDB) or black in case of missing information. Node sizes are proportional to normalized betweenness centrality and fitness for the host and pathogen proteins, respectively. Edge size and width correspond to the PPI ranked score (0-1 scale). Some of the PPIs with the highest final scores have been highlighted with colors to show how the highest-ranked PPIs from our predictive algorithm allow reconstructing of a highly connected subnetwork. The network was designed with Cytoscape.

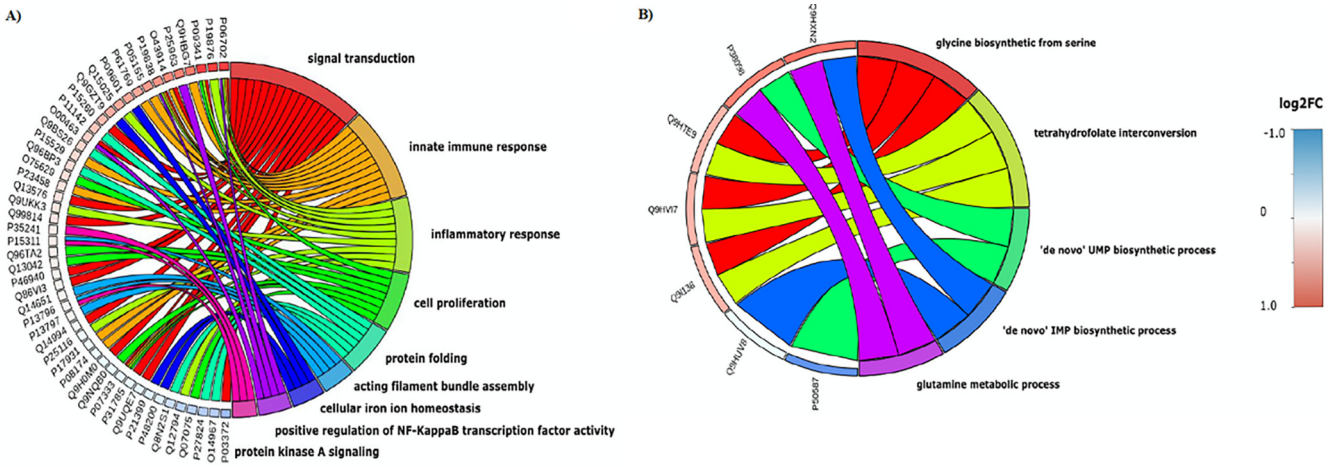

Fig. 6. Gene Ontology enrichment analysis of proteins involved in the highest scoring PPIs. Biological processes (BP) are depicted for (A) host and (B) pathogen proteins. Log fold changes for individual proteins are assigned from sequence similarity to log2 fold changes from DualSeqDB. Gene ontology analysis was performed using David and chord diagrams were drawn using the *circlize* package in R.

and nucleotide biosynthesis, as well as folate biosynthesis, all of them required for bacterial proliferation. These biosynthetic routes have been used as molecular targets for the development of antimicrobials, e.g., the antibiotic trimethoprim is a dihydrofolate reductase inhibitor.

3.3. Benchmarking with interolog prediction servers

We further compared HPIPred with two publicly available predictive software called BIPS (Biana Interolog Prediction Server) [9] and PredHPI [10]. Both algorithms predict putative PPIs based on interolog information. We used *Homo sapiens* and *Pseudomonas aeruginosa* PAO1 proteomes as query inputs to compare the results to our consensus interactome (Supplementary Information, Table S2). In the case of PredHPI, we recovered 37.815 interactions using default parameters (30 % similarity). Surprisingly, some proteins had many interactions, such as immunoglobulin IGHV4-31, efflux pump MacB and plasmin, with 8.730, 1.972, and 1.472 interactions, respectively. In fact, 10 proteins account for 50 % of the total number of interactions. These numbers are disproportionate and may represent a bias towards overrepresented entries in the positive databases used to train the predictor. Even increasing the similarity score to 40 %, we recovered 13.941 interactions, with 3.388, 177, and 156 interactions for IGHV4-31, MacB, and plasmin, respectively. Again, a single protein represents ~ 25 % of the total number of interactions, suggesting a high bias in the predicted interactome. Conversely, in BIPS, we obtained very few predictions using default parameters (<20), probably because the sequence similarity threshold was very restrictive (80 %). Hence, we relaxed the filtering criteria involving identity similarity to 40 %. BIPS predictive tool generated an interactome consisting of 963 PPIs, a more manageable set with a size similar to our predicted interactome. The results from BIPS and our algorithm were then compared by creating an intersection of the predicted interactomes, which revealed that both methods shared a total of 262 common PPIs. We represented these common PPIs as a network in Cytoscape (Fig. 7) and found that, in general, the shared PPIs maintained a certain degree of network connectivity and the proteins involved presented a high score in terms of betweenness centrality and fitness for the host and the pathogen, respectively. These results suggest that controlling for false positives is essential for useful predictions. Otherwise, a lot of time and resources would be wasted on experimental validation.

4. Discussion

Infectious diseases are a growing health concern worldwide, specially due to the increase in multi-drug-resistant bacteria [30]. These pathogenic bacteria cause prolonged hospitalizations, higher costs for medical treatment, and increased mortality rates. In this sense, protein interactions between pathogenic bacteria and their natural hosts play a key role in the infection mechanism and a thorough understanding of their complex interplay [31] is required for the development of new antibiotics. Numerous experimental techniques, such as yeast two-hybrid, pulldown assays, or co-immunoprecipitation, are currently used for the detection of these interactions, which are collected in databases through literature mining or manual curation. However, experimental techniques are time-consuming, costly, and suffer from low specificity [32], making it unfeasible to evaluate all possible protein–protein interactions.

Recently, computational approaches such as machine learning, homology-based methods, or structure-based methods, have allowed the prediction of putative PPIs, complementing experimental techniques [33]. The main limitation of these prediction methods lies in the high rate of false positives, mainly due to the lack of robust databases of experimentally validated PPIs [16]. This data shortage causes the prediction methods to perform poorly in terms of generalization. HPIPred has been developed as a tool that could help to reduce the false positive rate compared to other methods. To this end, HPIPred predicts putative PPIs through the

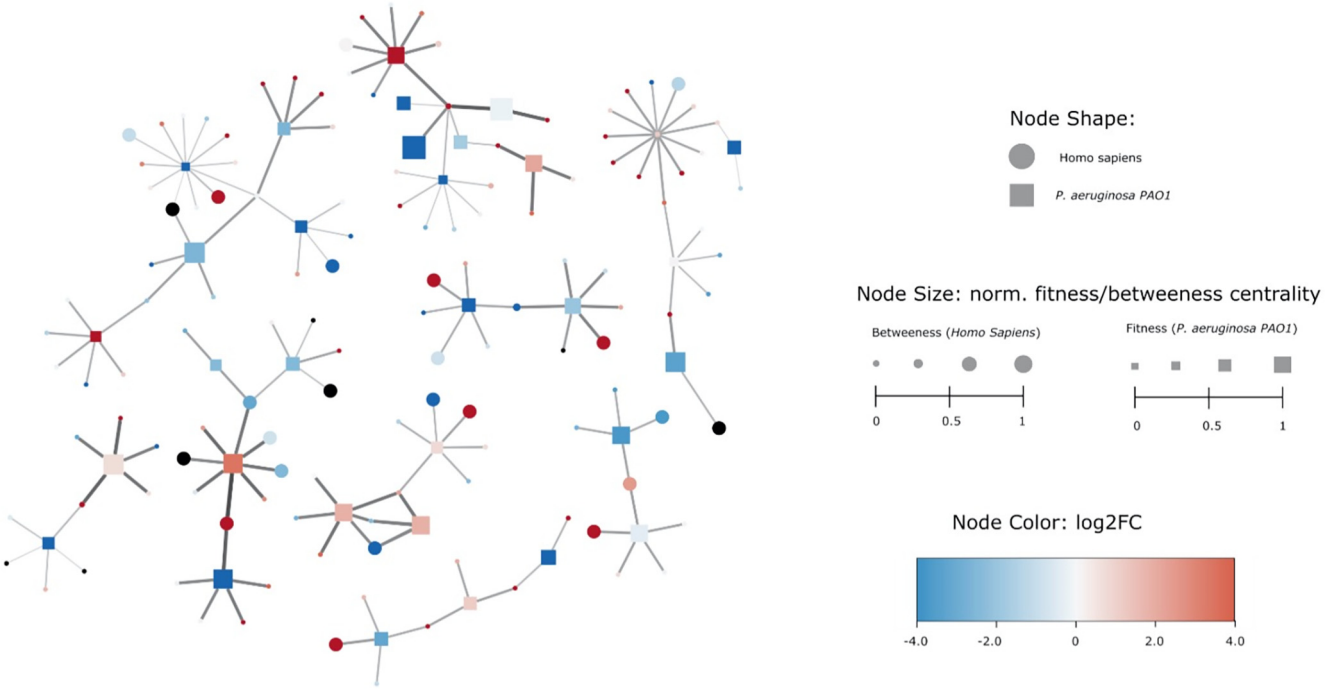

Fig. 7. Network representation of the common PPIs by BIPS and HPIPred. Host and pathogen proteins are represented by circles and squares, respectively. Nodes are colored according to the normalized expression changes (computationally derived from DualSeqDB) or black in case of missing information. Node sizes are proportional to normalized betweenness centrality and fitness for the host and pathogen proteins, respectively. Edge size and width correspond to the PPI ranked score (0-1 scale). The network was designed with Cytoscape.

numerical encoding of proteins based on physicochemical properties of the amino acids and integrates these predictions with biologically relevant data. Such data include information on the *in vivo* relevance of bacterial genes to the infection process, gene expression changes *in vitro* and *in vivo*, as well as topology information that allows highlighting the importance of central hubs. By using the *Homo sapiens* and the *Pseudomonas aeruginosa* PAO1 proteomes as input to our prediction tool, we generated 763 host-pathogen interactions displaying a highly connected network. We expect that our prediction tool will provide a more realistic picture of host-pathogen interactomes and help pave the way for the prioritization of PPIs that can be explored as potential targets for the development of new antibacterial drugs.

CRediT authorship contribution statement

Javier Macho Rendón: Data curation, Formal analysis, Methodology, Software, Writing – original draft. **Rocio Rebollido-Ríos:** Data curation, Formal analysis, Methodology, Writing – review & editing. **Marc Torrent Burgas:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by a Research Grant 2022 of the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) and the Spanish Ministerio de Ciencia e Innovación (SAF2017-82158-R, PDC2021-121544-I00 funded by MCIN/AEI /10.13039/501100011033 and European Union Next GenerationEU/ PRTR, and project PID2020-114627RB-I00 funded by MCIN/AEI /10.13039/501100011033), all to MT. R.R.-R was a recipient of an INCOMED Marie Curie Fellowship at the time.

Author contributions

MT designed, directed, obtained funding for, and coordinated the study. JM and RR-R performed the experiments. JM wrote an initial version of the paper, subsequently edited by MT. All authors contributed to editing and revising the final version of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.026.

References

[1] Ribet D, Cossart P. How bacterial pathogens colonize their hosts and invade deeper tissues. Microbes Infect 2015;17:173–83.
[2] Crua Asensio N, Macho Rendon J, Torrent Burgas M. Time-resolved transcriptional profiling of epithelial cells infected by intracellular Acinetobacter baumannii. Microorganisms 2021;9.
[3] de Groot NS, Torrent Burgas M. Bacteria use structural imperfect mimicry to hijack the host interactome. PLoS Comput Biol 2020;16:e1008395.
[4] Cossar PJ, Lewis PJ, McCluskey A. Protein-protein interactions as antibiotic targets: a medicinal chemistry perspective. Med Res Rev 2020;40:469–94.
[5] Karimova G, Ladant D, Ullmann A. Two-hybrid systems and their usage in infection biology. Int J Med Microbiol 2002;292:17–25.
[6] Gagarinova A, Phanse S, Cygler M, Babu M. Insights from protein-protein interaction studies on bacterial pathogenesis. Expert Rev Proteomics 2017;14:779–97.
[7] Jean Beltran PM, Federspiel JD, Sheng X, Cristea IM. Proteomics and integrative omic approaches for understanding host-pathogen interactions and infectious diseases. Mol Syst Biol 2017;13:922.
[8] Gómez Borrego J, Torrent Burgas M. Analysis of Host-Bacteria Protein Interactions Reveals Conserved Domains and Motifs That Mediate Fundamental Infection Pathways. International Journal of Molecular Sciences 2022;23(19):11489.
[9] Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. Nucl Acids Res 2012;40:W147–51.
[10] Loaiza CD, Kaundal R. PredHPI: an integrated web server platform for the detection and visualization of host-pathogen interactions using sequence-based methods. Bioinformatics 2021;37:622–4.
[11] Das D, Krishnan SR, Bulusu G, Roy A. in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 935-938 (2019).
[12] Zhang SB, Tang QR. Protein-protein interaction inference based on semantic similarity of Gene Ontology terms. J Theor Biol 2016;401:30–7.
[13] Mariano R, Wuchty S. Structure-based prediction of host-pathogen protein interactions. Curr Opin Struct Biol 2017;44:119–24.
[14] Kaundal R, Loaiza CD, Duhan N, Flann N. deepHPI: a comprehensive deep learning platform for accurate prediction and visualization of host-pathogen protein-protein interactions. Brief Bioinform 2022;23.
[15] Li F, Zhu F, Ling X, Liu Q. protein interaction network reconstruction through ensemble deep learning with attention mechanism. Front Bioeng Biotechnol 2020;8:390.
[16] Dunham B, Ganapathiraju MK. Benchmark evaluation of protein-protein interaction prediction algorithms. Molecules 2021;27.
[17] Durmus Tekir, S. et al. PHISTO: pathogen-host interaction search tool. Bioinformatics 29, 1357-1358 (2013).
[18] Tiessen A, Perez-Rodriguez P, Delaye-Arredondo LJ. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. BMC Res Notes 2012;5:85.
[19] Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucl Acids Res 2000;28:374.
[20] Rendon JM, Lang B, Tartaglia GG, Burgas MT. BacFITBase: a database to assess the relevance of bacterial genes during host infection. Nucl Acids Res 2020;48:D511–6.
[21] Macho Rendon J, Lang B, Ramos Llorens M, Gaetano Tartaglia G, Torrent Burgas M. DualSeqDB: the host-pathogen dual RNA sequencing database for infection processes. Nucl Acids Res 2021;49:D687–93.
[22] Urban M et al. PHI-base: the pathogen-host interactions database. Nucl Acids Res 2020;48:D613–20.
[23] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001;411:41–2.
[24] Crua Asensio N, Munoz Giner E, de Groot NS, Torrent Burgas M. Centrality in the host-pathogen interactome is associated with pathogen fitness during infection. Nat Commun 2017;8:14092.
[25] Ashtiani M et al. A systematic survey of centrality measures for protein-protein interaction networks. BMC Syst Biol 2018;12:80.
[26] Szklarczyk D et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucl Acids Res 2021;49:D605–12.
[27] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. Complex Systems 2006;1695.
[28] Shannon P et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.
[29] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
[30] Antimicrobial Resistance C. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet 2022;399:629–55.
[31] de Groot NS, Torrent Burgas M. A coordinated response at the transcriptome and interactome level is required to ensure uropathogenic escherichia coli survival during bacteremia. Microorganisms 2019;7.
[32] Zhou M, Li Q, Wang R. Current experimental methods for characterizing protein-protein interactions. ChemMedChem 2016;11:738–56.
[33] Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. Chem Rev 2016;116:4884–909.