Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Structural Biology

Structure-based prediction of host–pathogen protein interactions

Rachelle Mariano[1] and Stefan Wuchty[2,3,4]

The discovery, validation, and characterization of protein-based interactions from different species are crucial for translational research regarding a variety of pathogens, ranging from the malaria parasite *Plasmodium falciparum* to HIV-1. Here, we review recent advances in the prediction of host–pathogen protein interfaces using structural information. In particular, we observe that current methods chiefly perform machine learning on sequence and domain information to produce large sets of candidate interactions that are further assessed and pruned to generate final, highly probable sets. Structure-based studies have also emphasized the electrostatic properties and evolutionary transformations of pathogenic interfaces, supplying crucial insight into antigenic determinants and the ways pathogens compete for host protein binding. Advancements in spectroscopic and crystallographic methods complement the aforementioned techniques, permitting the rigorous study of true positives at a molecular level. Together, these approaches illustrate how protein structure on a variety of levels functions coordinately and dynamically to achieve host takeover.

Addresses
[1] Brigham & Women's Hospital, Harvard Medical School, Harvard University, Cambridge, MA, United States
[2] Dept. of Computer Science, Univ. of Miami, Coral Gables, FL, United States
[3] Center for Computational Science, Univ. of Miami, Coral Gables, FL, United States
[4] Sylvester Cancer Center, Univ. of Miami, Miami, FL, United States

Corresponding author: Mariano, Rachelle (rmariano@g.harvard.edu)

Current Opinion in Structural Biology 2017, **44**:119–124

This review comes from a themed issue on **Sequences and topology**

Edited by **Ramanathan Sowdhamini** and **Kenji Mizuguchi**

http://dx.doi.org/10.1016/j.sbi.2017.02.007

0959-440X/© 2017 Elsevier Ltd. All rights reserved.

The accurate determination of protein–protein interactions between bacterial, viral, and parasitic pathogens and their human hosts harbors great medicinal potential, as these discoveries could be used to target specific disease-related interfaces with minimal disruption of the underlying human interaction network. Indeed, elucidating host–pathogen protein–protein interactions (HP-PPIs) for therapeutics drives their intensive computational and experimental study, and rapidly improving approaches have generated valuable high-fidelity HP-PPI candidates. Benchside high-throughput methods often bear a considerable proportion of false positives when applied to HP-PPI prediction. Moreover, exogenous expression of pathogenic proteins remains difficult, and most results must be translated across evolutionarily distant species. Computational inference of HP-PPIs can identify small subsets of highly probable interactions for informed experimental follow-up by techniques such as nuclear magnetic resonance microscopy (NMR) and crystallography. Combined, these methods allow researchers to not only ascertain how a pathogenic protein interacts with its host on a molecular scale, but also how such interactions function in a larger cellular network.

Computational HP-PPI prediction based on sequence and domain information: homology-based approaches

While first limited to intraspecies interactions, sequence similarity-based approaches have since extended to interspecies PPI prediction. Since high primary sequence similarity implies an interaction – an interolog – these methods map known interaction interface sequences onto homologous or orthologous pairs of sequences in different organisms. For example, such sequence comparisons yielded HP-PPIs between *Plasmodium falciparum* (*P. falciparum*) [1–3] and *Helicobacter pylori* (*H. pylori*) [4] and their human host. Interolog screens benefit from their straightforward execution as well as abundant protein sequence information from which to mine data [5]. As the interologous proteins should demonstrate at least 80% sequence similarity, the ability to correctly determine HP-PPIs from interologs rapidly decreases with evolutionary distance. Additionally, pathogens are locked in a biological 'arms race' with their hosts, and their proteins may experience rapid changes in sequence that affect the fidelity of interolog screens [6••]. Interolog screens also have a penchant to generate a huge amount of false positive hits. Therefore, further computational investigation of potential hits involves filtering based on the cellular localization, biological functions, and expression profiles of putative HP-PPIs to significantly improve the quality of potential HP-PPI candidates [1–4].

Machine learning approaches using derived sequence-based features have also procured possible HP-PPIs. Shen *et al.* [7] represented sequences of interacting proteins as numerical profiles of the occurrence of amino-acid triplets. To predict interactions between human proteins, they utilized support vector machine algorithms (SVM) that were trained by carefully picked positive and negative training sets of protein interactions. This approach was similarly applied to interactions between human proteins and the malaria parasite *P. falciparum* [8]. Such a representation reduces the size of the feature space but may impair the quality of results. All approaches that predict HP-PPIs via supervised machine learning-whether with sequence or higher order information- need appropriate positive and negative training sets to robustly classify interacting proteins. Yu *et al.* [9] showed that the choice of non-interactions in the training data greatly impacts the accurate identification of interacting vs. non-interacting pairs. Tools balancing negative example selection have been recently developed to combat this issue. In particular, Eid *et al.* [10] developed a dissimilarity-random-sampling algorithm for the determination of unlikely occurring interactions between human host and pathogen proteins. The authors sampled highly dissimilar protein sequences from other viruses compared to interacting proteins of the virus in question to generate a negative training set. These training sets trained a SVM with prediction accuracies up to 86% [10], suggesting that the skilled choice of negative training sets drives the reliability of predicted HP-PPIs.

Domain-based approaches

Computational inference of HP-PPIs often combines primary sequence similarity with higher order structural information from motifs and domains to increase prediction accuracy [11]. A protein domain is usually defined as a conserved part of a protein's sequence and three-dimensional structure that mediates the protein's biological functions while folding and evolving independently [12]. Since domain–domain interactions (DDIs) are largely considered to drive PPIs, numerous studies have used known intra-species DDIs as a basis for the prediction of HP-PPIs. Furthermore, unbiased approaches to elucidate significant predictive features of HP-PPIs have repeatedly emphasized their role [9]. In particular, out of 44 descriptors involving amino acid frequencies of host and pathogen sequences, protein–domain associations appeared to have the highest predictive effect when used with SVM and random forest (RF) algorithms [13]. In a different study, DDIs were included in an 18-dimensional vector and combined with topological sequence and functional characteristics to predict interactions between proteins of HIV-1 using different variations of neural network methods [14] that outperformed RFs [15].

Combining domain-based data with primary sequence homology searches of interacting domains allowed the large-scale detection of hypothetical interactions between proteins of *H. pylori*, HIV-1, and *Salmonella* with their human host [4]. DDIs have also been combined with protein sequence *k*-mers and topological properties of host proteins in a human protein interaction network to predict host–pathogen interactions using a SVM [16••]. Notably, the use of DDIs allowed prioritization of proteins with extracellular or trans-membrane domains to assess interactions driving invasion and intracellular signaling [4,16••]. Domain-based prediction also assisted in the identification of common functional features that allow pathogens to interact with more than one host [17].

DDIs can also be employed separately from primary sequences to derive HP-PPIs. Itzhaki *et al.* integrated sets of protein interactions from various organisms with verified protein–domain profiles, assuming that intraspecies DDIs similarly connect in HP-PPIs [18]. The authors modeled the probability that proteins with certain domains interact in a Bayesian framework and generated a protein interaction network between *P. falciparum* and the human host. Remarkably, interactions thus obtained featured significant co-expression of involved pathogen and human proteins, illustrating the high probability of interaction that DDI-based HP-PPIs can achieve. Liu *et al.* [19] used an expectation maximization algorithm to find expression-correlated interactions between *P. falciparum* and human erythrocytes that were subsequently verified using expression data.

Motif and integration-based approaches

Motif–domain and motif–motif interactions have also gained traction as foundations for HP-PPI prediction, as short linear motifs have proved vital for host–pathogen protein binding. Evans *et al.* annotated short eukaryotic linear motifs (ELMs) in HIV-1 proteins and used human counter domains that interact with these ELMs to generate an HIV-1 and human interactome [20]. Segura-Cabrera *et al.* [21] combined motif information from 3-D interaction databases with stringent filters to create an infectome representing HP-PPIs of 5 viruses with the human host, integrating surface accessibility and structural information. Although these studies predicted HIV-1 HP-PPIs with similar techniques, their results differed as a consequence of discrepancies in filtering and motif definition. Although stringent filters were applied to secure biologically meaningful results, both studies provided a plethora of interactions [21]. This is a standing issue in computational HP-PPI prediction, as results depend on preferred methods and suffer from persistence of false positive hits.

Although primary and secondary sequence integration enhances computational HP-PPI predictions, auxiliary data are increasingly used to curb the impact of false positives. In particular, current approaches are assimilating domain, sequence, and ELM data with gene ontology (GO) features, graph topological properties, and gene co-expression data to train HP-PPI classifiers, instead