Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

Computational biology and machine learning approaches to study mechanistic microbiome-host interactions

Padhmanand Sudhakar[1,2,3], Kathleen Machiels[1], Séverine Vermeire[1,4]

Affiliations

[1]KU Leuven Department of Chronic Diseases, Metabolism and Ageing, Translational Research Center for Gastrointestinal Disorders (TARGID), Leuven, Belgium.

[2]Earlham Institute, Norwich, United Kingdom.

[3]Quadram Institute, Norwich, United Kingdom.

[4]Department of Gastroenterology and Hepatology, KU Leuven University Hospitals Leuven, Leuven, Belgium

1. Abstract

The microbiome, by virtue of its interactions with the host, is implicated in various host functions including its influence on inflammation, nutrition, and homeostasis. Although driven by a complex combination of intrinsic and extrinsic factors, many chronic diseases such as diabetes, cancer, Inflammatory Bowel Disease among others are characterized by a disruption of microbial communities in at least one biological niche/organ system. Various molecular mechanisms between microbial and host components such as proteins, RNAs, metabolites etc have recently been elucidated, thus filling many gaps in our understanding of how the microbiome modulates host processes. Concurrently, high throughput technologies have enabled the profiling of heterogeneous datasets capturing community level changes in the microbiome as well as the host responses. However, due to pragmatic limitations with respect to parallel sampling and analytical procedures, big gaps still exist in terms of how the microbiome mechanistically influences host functions at a systems and community level. In the past decade, various computational biology and machine learning methodologies and approaches have been developed with an aim to fill these existing gaps. Due to the agnostic nature of the tools, they have been applied in various disease contexts to analyze and infer the interactions between the microbiome and host molecular components, and in the case of a few selected tools, on downstream host processes. Generally, most of the tools are enabled by frameworks to statistically or mechanistically integrate different types of -omic and meta -omic datasets followed by functional/biological interpretation. In this review, we provide an overview of the landscape of computational approaches for investigating mechanistic microbiome-host interactions and their potential benefit for basic and clinical research. These could include but are not limited to the development of activity and mechanism based biomarkers, uncovering mechanisms for therapeutic interventions and generating integrated signatures to stratify patients.

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

2. Introduction : microbiome-host interactions

Across different niches and ecosystems, micro-organisms including bacteria, viruses, archaea inhabit a wide range of hosts[1]. This community of microbes imparts various functions such as making nutrients accessible to the host[2], modulating the host immune system[3], warding off pathogens[4], maintaining homeostasis[5,6] among others. These functions are in turn driven primarily by inter-species molecular interactions between microbial and host molecules such as proteins, RNA and metabolites[1,7]. Deciphering these interactions could not only help us understand the mechanisms which underlie the microbe-host cross-talk but also provide us with some insights into formulating therapeutic strategies aimed at maintaining health and/or ameliorating disease states. The past couple of decades have witnessed a surge in research interest to study microbial communities (and their interactions) which inhabit various niches - from the gut to the soil ecosystem. This was made possible by technological advancements leading to plummeting costs of 16S and metagenomic sequencing, higher sequencing depth and resolution[8–10], novel in-vitro systems[11–13] and new methodologies for high-throughput profiling of multiple -omic data types such as metaproteomics, metabolomics, lipidomics etc[14,15]. However, due to many other limitations related to scale, scope, feasibility and sample availability for parallel omic read -outs, experimentally determining the inter-species microbe-host interactions is a challenging task[16]. Computational methods can overcome some of these limitations and hence has thrown open the door for enhancing our understanding of microbe-host interactions[17]. In this review, we outline some key concepts, tools and methods involved in computationally inferring the molecular mechanisms mediating microbe-host interactions.

3. Computational methods in microbiome-host interactions : filling the gaps

To partially overcome the challenges and gaps in experimentally verifying the molecular mechanisms involved in microbe-host interactions, computational methods have been adopted and modified to this end. Such methods bring in various advantages to the analysis of microbe-host interactions. These include their attributes of (a) enhancing scalability, i.e, perform the computational inferences for a large number of variables and samples (b) improving reproducibility (c) assessing performance by using a series of metrics (d) shortlisting and prioritizing interactions (e) and thereby enabling the finetuning of hypothesis for experimental and/or epidemiological studies.

4. Classification of methods

From a molecular mechanistic view-point, the most widely studied interaction types in interspecies cross-talks include (a) protein-protein interactions (PPIs), (b) RNA-mediated interactions and (c) microbe-host metabolic networks. Accordingly, many of the computational methods developed to investigate microbe-host interactions have focussed on the three above-mentioned interaction types (**Figure 1**). As a fourth method category, integrated pipelines come into the picture and combine multiple microbial and host -omic data types and networks to infer the cumulative functional effects of inter-species interactions/communication on the host.

4.1 Approaches inferring mechanistic metabolic interactions

The metabolomic layer and the interactions within have a prominent influence on both health and disease states associated with alterations in microbiota composition[18,19]. Metabolic networks can thus represent and capture the underlying mechanisms driving various

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

phenotypes[20–22]. Computational approaches aimed at inferring the microbe-host co-metabolic networks in the literature can be classified into three prominent categories namely (a) Community wide metabolic network modelling using metagenomic datasets (b) High throughput data driven approaches using metabolic and metagenomic datasets and (c) Genome scale reconstruction applying constraint-based modelling approaches. The first two methods do not provide any mechanistic insights and hence won't be surveyed in this review.

Genome scale reconstruction models[23,24] on the other hand provide organism resolved mechanistic information by integrating multiple inputs. These inputs include the curated genome scale metabolic models of both the host and microbial species, high-throughput meta -omic datasets including those of metabolites, reaction fluxes, biochemical traits and accessory phenotypic data[23,24]. However, due to the strenuous nature of various steps involved in constructing the models and/or in scaling it up to multiple species or multiple hosts, only a handful of studies have applied this concept to infer microbe-host co-metabolic interactions (**Table 1**). The reported studies have been distributed across many different ecological contexts such as the human and rumen gut ecosystems[25], microbe-plant interactions, human alveolar macrophages, the effect of viral demands on the metabolism of human macrophages, microbe-host interactions in Parkinson's Disease to name a few. By incorporating the individual reconstructed metabolic models of tomato (*S. lycopersicum*) and the tomato late blight pathogen *P. infestans*, Rodenburg et al pointed out specific pathways which mediate the dependencies of the pathogen on the metabolism of *S. lycopersicum*[26]. Furthermore, by over-laying dual RNA-seq transcriptomic datasets from the host-pathogen duo into the co-metabolic network, various metabolic changes characterizing the scavenging nature of *P. infestans* were revealed. A similar study was performed in a mammalian setting wherein co-metabolic interactions and metabolic exchanges were inferred between the respiratory pathogen *M. tuberculosis* and human alveolar macrophages[27]. Unsurprisingly, given the advancement in terms of data generated and metabolic models made available, most of the genome-scale metabolic reconstruction studies (**Table 1**) were carried out for the gut ecosystem[25,28,29,30]. A representative study of the gut ecosystem integrated two previously published constraint-based models of mouse and a commensal gut bacterium *B. thetaiotaomicron*[29]. The integrated metabolic model could capture many of the phenotypes exhibited *in vivo* namely the dependence of *B. thetaiotaomicron* on glycans derived from the metabolism of the host as well as the host diet itself[29]. Due to the mechanistic nature of such models, they can be used as a template for further integrating other -omic datasets[23,24]. This not only refines the models thereby increasing their predictive power but also assigns contextuality.

4.2 Approaches inferring protein-protein interactions (PPIs)

PPIs are one of the most well-studied interaction types mediating inter-species communication[31]. Accordingly, a large number of computational microbe-host interaction studies have focussed on PPIs. Congruently, PPI-based approaches have also been propelled by the adoption of concepts from other domains of computational biology and computational sciences in general. Hence, PPI-based approaches can be sub-classified into four predominant methods (**Table 2**) depending on the concepts used (1) Machine learning based PPI methods (2) Structural feature based PPI methods (3) Data/Literature mining based PPI methods (4) Interolog based PPI methods. In this section, we provide a brief overview of the concepts involved in each of these methods (**Table 2**) and provide a few representative examples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

4.2.1 Structural feature based PPI methods

Interactions between proteins are usually a by-product of physical interactions between structural features of the proteins and/or could be characterized indirectly by co-occurring functional features of the proteins[32]. Structural features of the proteins include their domain and motif architectures/compositions, amino acid composition and frequencies, post-translational modification signatures, amino acid k-mers, mimicry motifs and 3D structural properties[32]. Structural feature based PPI prediction, applied initially for intra-species PPIs, was subsequently extended to inter-species studies. Essentially, the fundamental principle on which structural feature based PPI prediction methods work involves the use of mechanistic evidence between structural features to identify potentially interacting proteins. Mechanistic evidence involving structural features include interactions between domains, between domains and motifs, post-translational modifications, pairwise structural similarity to name a few[32]. Such structural studies have been confined to considerably well studied species pairs involving *H. sapiens* and prominent viral and bacterial pathogens (**Table 2**). Along with pairwise structural similarity-based methods using 3D protein complexes, domain-domain interaction (DDI) and domain-motif interaction (DMI) based methods are one of the most commonly used methods within the structural feature based methodological framework for predicting inter-species PPIs. Due to the ease of annotating domains and motifs, DDI- and DMI-based methods have been harnessed widely (**Table 2**). While DDI based methods have been applied to infer PPIs for a large number of species-pairs including Human–*P. falciparum*[33], Human–*M. tuberculosis*[34,35], Human–*L. interrogans*[36], Human–*L. biflexa*[36], Human–papillomavirus type 16[37], Arabidopsis–*P.syringae*[38], Rice–*X. oryzae*[39], they have the inherent disadvantage of not being able to explicitly discern directionality.

On the other hand, DMIs provide directionality for PPIs thus indicating the flow of signal transduction[40,41]. For example, if a microbial protein A contains a domain known to be interacting with a motif on the host protein B, it is graphically represented as A➡B, translating into "microbial protein A modulates host protein B". Due to their specificity, DMI-based methods are preferred over DDI based methods for research questions seeking to answer the role of post-translational modifications elicited on host proteins by microbial proteins or vice versa. However, due to the short sequence length of protein sequence motifs, even the most stringent search strategies have the tendency to result in thousands of false-positive hits while performing motif search on a proteome-wide basis[42,43]. Therefore, proper quality controls need to be applied to filter out false-positives based on structural properties such as the occurrence of truly interacting motifs within disordered regions and outside globular domains[42,43].

Several studies (**Table 2**) have been conducted to apply the principles of DMIs to predict PPIs for multiple microbe-host species-combinations including grass carp-grass carp reovirus[44], human-multiple bacterial pathogens[45], human-multiple viruses[46] and human-HIV[47]. By integrating DMI predictions between grass carp and grass carp reovirus (GCRV) proteins with differential gene expression and tissue-specific gene expression followed by functional enrichment, Zhang et al[44] were able to pinpoint several signalling pathways modulated by GCRV. The authors also highlight an enrichment of host genes expressed in the intestinal niche suggesting that GCRV might have a higher influence on the gut. Recently, we conducted a study[45] using DDI and DMI based methods to identify cross-talks between several bacterial pathogens including *Salmonella* and autophagy – a prominent biological process involved in host cellular homeostasis. Firstly, to identify microbial proteins targeted by selective autophagy, we scanned the bacterial proteins for the presence of the recognition motifs corresponding to the selective autophagy receptors p62 and NDP52 and the autophagy adapter

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

protein LC3. Conversely, to infer the modulation of host autophagy by the bacterial pathogens, DMI and DDI based methods were used to identify the bacterial proteins which are able to bind to/modulate the 37 core autophagy host proteins. By overlapping the two above-mentioned sets of predictions, bacterial proteins involved in interplays were identified. Such bacterial proteins not only modulate host autophagy but are also targeted by the host autophagy machinery for clearance and degradation. This was followed by experimentally verifying the effect on autophagy of a *Salmonella* protease involved in human-*Salmonella* interplay.

A variation of the motif-based methodologies is the use of motifs to characterize pathogen mimicry. This essentially involves the identification of eukaryotic linear motifs on microbial proteins which in turn can hijack host proteins and thereby promote antagonistic binding[48,49]. Motif-mediated molecular mimicry therefore rewires the host signalling and regulatory networks by titrating essential host proteins and enabling the microbe to create favourable micro-environments in the host cell by altering immune responses for example[49,50]. In addition to motifs, molecular mimicry can also be mediated at the level of protein, structural and interface levels. At the protein level, specific studies investigating the role of molecular mimicry in the pathogenesis of prominent bacterial pathogens[51] including *S. typhimurium* and *Human respiratory syncytial virus*[52] have been carried out (**Table 2**). At the interface level, Guven-Maiorov et al 2017[53] devised a computational method to infer mimicry induced by a prominent gastric cancer causing pathogen *H. pylori*. Besides DDI and DMI based methods, researchers have also used other structure-based methodologies such as pairwise structural similarity (PSS) to predict inter-species PPIs. PSS methods at their very core are based on the premise that proteins possessing similar structures have a greater probability of interacting with the same set of protein partners[32] and have been applied to infer the interactions with the host of various pathogens such as Dengue virus[54,55], HIV[56], *M. tuberculosis*[56], West Nile virus[55], Chandipura virus[57] and other viral pathogens[58,59].

As a means of ensuring proper quantitative evaluation of de-novo PPI predictions, emerging computational methods such as machine learning (ML) have been used in conjunction with structural-feature based PPI prediction methods. In order to avoid repetitions, methods using ML for evaluating the performance of structural feature dependent PPI predictions are discussed in the next section.

4.2.2 Machine learning based PPI methods

Due to their ability to discern complex patterns among a large number of features in big datasets, machine learning (ML) methods have found favour in various applications of computational biology and bioinformatics[60] including the prediction of microbe-host molecular interactions. A variety of supervised and unsupervised methods have been used to predict the interactions between microbial and host proteins (**Table 2**). In general, supervised machine learning methods predicting inter-species PPIs utilize features from "gold-standard" interaction datasets to identify potential protein-protein interaction pairs from the user provided list of microbial and host proteins[61]. In supervised methods, the "gold-standard" datasets are either compiled from high-throughput experimental methodologies or from curated lists of interactions from the literature[61]. In the case of ML being used in combination with "interolog" based methods (explained in section 4.2.5), "gold-standard" PPI datasets can also be retrieved from other related or unrelated microbe-host species pairs depending on the scope of the study. Some of the features used to infer de-novo PPI predictions include protein properties such as post-translational modifications, chemical composition, tissue distribution, molecular weight, domain/motif compositions, ontologies, gene expression, amino-acid frequencies, homology

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

to human binding partners, relevance of proteins in host network etc to name a few. By using these features, supervised methods are able to discern truly interacting protein pairs from all possible pairs of microbial and host proteins[61].

Supervised methods can also be differentiated by the kind of ML methodology / model used for the task of rightly classifying truly interacting protein pairs. Several supervised studies employing individual ML models (such as I2-regularized logistic regression[62], SVM[63–65], RF[66] etc) or a combination of different models (usually known as ensemble learning) have been applied to infer PPIs between microbial and host species. For instance, using four different ML models namely Random Forests (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN) and K-Nearest Neighbors (K-NN), and multiple lines of -omic evidence including PPIs as predictive features, Leite and colleagues devised a supervised protocol to accurately predict bacterium-phage interactions[67]. The model, due to its generic nature, can also be used to predict interactions between any two given species, given the availability of informative feature sets. Ensemble learning has also been used to predict PPI based HIV-human and hepatitis C virus-human networks[68,69]. Various auxiliary algorithms have been used in conjunction with machine learning methods to predict inter-species PPIs. An example of such a study includes the use of a novel protein sequence based feature extraction method called Location Based Encoding with different classifier models including RFs to predict interactions between proteins from two important pathogens - *B. anthracis* and *Y. pestis* and human proteins[66].

Supervised methods are sometimes faced with the small size of "gold-standard" datasets which restricts the inference and prediction of proteome wide PPIs between the full list of proteins of any two given species. Mei and Zhu harness the power of Multi-instance AdaBoost, which is a multi-instance learning based ML method, to reconstruct proteome-wide Human T-cell leukaemia virus-human PPI networks using homology knowledge derived protein features[70]. The dearth of true interacting protein-pairs has also prompted researchers to use unsupervised or semi-supervised approaches to infer microbe-host PPIs. Qi et al complement the list of true interactions with a list of protein-pairs wherein association evidence exists with no interaction evidence between the proteins of a pair[71]. Supervised learning is achieved with a multilayer perceptron network and by using the true interaction list. Subsequently, the semi-supervised approach uses the same network layers of the supervised classifier but instead trains on the protein-pairs with association evidence only. By using this hybrid approach, the authors report improved performance for predicting interactions between HIV and human proteins[71].

4.2.3 Data/Literature mining based PPI methods

Eventhough many databases have been compiled to collect, curate and store microbe-host PPIs[72–76], these are either confined to well-studied pathogens and are predominantly comprised of interactions from high-throughput experiments. Contrastingly, in the literature, there exist inter-species PPIs from low-throughput experiments with some of them from non-model organisms, and commensal microbes, mostly reported distributed over several individual studies. Very often, the inter-species PPI databases and repositories do not capture these sparse interactions. Hence, researchers have adapted and modified literature and data mining tools to search for and extract microbe-host PPIs from existing literature. Retrieving such PPIs not only helps in increasing the number of true positive and true negative interactions (which helps aid the predictive performance of algorithms) but also extends our knowledge of existing microbe-host interactions. Motivated by the above explained need to mine-out microbe-host PPIs, Thieu et al[77] combine and compare the performance of a language based method based on a link

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

grammar parser to a supervised ML methodology (SVM) and report that the combined approach results in a higher classification accuracy when compared to existing literature mining methods. As part of a bigger analytical framework aimed at uncovering the cellular mechanisms involved in human B lymphocytes during Epstein-Barr virus infection, Li et al[78] use a big-data mining methodology to identify a diverse range of inter-species molecular interactions including PPIs. Similar text/data mining approaches were executed to extract PPI-mediated interactions of the human host with multiple viruses such as Hepatitis C virus[79], Influenza A virus[80] and HIV[81,82,83] (**Table 2**).

4.2.4 Interolog based PPI methods

For most species-pairs of interest, especially those belonging to the category of non-model organisms, there is a scarcity of experimentally verified PPIs. This has necessitated the development of novel bioinformatic methods, one of which is the inference of interactions from existing experimentally determined inter-species PPIs[84]. These types of methodologies are usually based on the principle of homology (hence the term "interolog" : meaning interacting orthologs) – either at the level of proteins or protein structural features or both. Protein structures used for homology based extrapolation include but are not limited to domains, motifs, amino-acid k-mers, and 3D structural properties[84]. Interolog based approaches have been applied to harness the large volume of experimentally verified PPI for model organisms including prominent bacterial/viral pathogens. Despite the potentially large coverage which can be achieved by such approaches, there exist several disadvantages of using interolog approaches as a silver bullet for inferring inter-species PPIs especially for novel species-pairs. These disadvantages are attributed to different pathogenic mechanisms and factors between the microbes in the context of infecting different host species, different cellular localizations for the orthologous microbial proteins, varying activity levels (expression, post-translational modifications etc) of the orthologous microbial proteins to name a few. Such differences lead to accessibility bottlenecks i.e ability of the proteins to physically access host proteins and thereby interact. Hence, interolog based approaches need to be complemented with additional filtering and quality control steps such as selecting proteins from infection-relevant cellular compartments, expression/activity measurements etc.

Interolog based methods have been used to infer inter-species PPIs for many prominent pathogens and parasites (**Table 2**). Different versions of the interolog approach have been used to extrapolate PPIs corresponding to interactions between the human host and various pathogens such as *P. falciparum*[85,86], *E. coli*[87], *S. typhimurium*[87,88], *Y. pestis*[87], *H. pylori*[89], HIV[56], *M. tuberculosis*[56,90], *C. burnetti*[91], *C. pseudotuberculosis*[92], *C. diphtheriae*[92] *and C. ulcerans*[92]. Using PPIs from the STRING database as the starting interaction set, Cuesta-Astroz et al 2019[93] used the interolog methodology to predict PPIs between 15 different eukaryotic pathogens and the human host. To assign species-specific and lifecycle- specific contextuality, the authors confined the analysis to proteins from particular cellular compartments which are relevant to the infection process. From the analysis of the ensuing PPI networks, various invasion and evasion mechanisms adopted commonly and specifically by particular parasites were inferred[93]. Schleker et al 2012[88] present another version of the interolog approach to predict human-*Salmonella* and *A. thaliana*- *Salmonella* PPI networks. As a source of template PPIs, publicly available interaction databases are used along with databases containing 3D structures between Pfam domains. As an add-on to the sequence based orthology of proteins, domain based orthology is also performed in order to reduce the false positive rates. Several additional filtering strategies such as restriction to predicted transmembrane proteins, relevance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

in host network and functional attributes such as gene ontology are used to make the PPIs more specific.

4.3 Approaches inferring RNA mediated interactions

The role of RNAs, especially non-coding RNAs such as long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) in mediating molecular microbe-host interactions have been reported in the literature[84,94,95]. RNA molecules are either secreted by the microbial cell into the host cell or are packaged into vesicles along with other molecules which are then taken up by the host cell by endocytosis[96–98]. Such microbial RNAs then modulate host cell activity by either binding to DNA, messenger RNAs or proteins. Thus, by salvaging and titrating host components, microbial RNAs modulate regulatory and signalling network and subsequently host cell activity[95,99,100]. However, in contrast to PPI based methods, even though RNA-mediated microbe-host interactions are well studied from an experimental point of view, very few methods or studies exist which have systemically and systematically applied computational analysis (**Table 3**). As such, the resources which exist in the domain of RNA-mediated microbe-host interactions comprise of databases such as ViRBase[94] which is predominantly a source of experimentally verified virus–host non-coding RNA-associated interactions. In addition, it also contains predicted binding sites of virus non-coding RNAs on host proteins and RNAs. The only study which comprehensively examines and evaluates the role of RNAs in microbe-host interactions is that of Demirci and Adan 2020[101] who investigated the roles in infection of miRNA-like sequences encoded within the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) genome. They used a modified version of izMiR[102], a SVM based ML method to predict miRNAs which are homologous to the human precursor miRNAs from miRbase. As a next step, the psRNATarget tool[103] was used to predict de-novo the human genes targeted by the inferred viral miRNAs. Functional analysis of the human genes targeted revealed that the viral miRNAs can affect various host processes including transcription, defense systems, Wnt and EGFR signalling pathways to name a few.

4.4 Approaches utilizing pipelines integrating multiple -omic datasets

Besides the computational methods based on particular types of molecular interactions, some integrated pipelines (**Table 4**) have been compiled to infer mechanistic microbe-host interactions. In general, such pipelines incorporate the prediction of at least one interaction type between microbial and host molecular components followed by various other functionalities such as integration of host responses. KBase[104] is an integrated bioinformatics platform enabling users to share datasets with the research community as well as facilitating the integration, and analysis of -omic datasets from microbes and plants by creating computational workflows. Recently, we developed MicrobioLink[105], an integrated pipeline which carries out de-novo DDI and DMI based microbe-host PPI prediction followed by quality control using information from disordered region predictions from built-in tools such as IUPred[106]. The pipeline then utilizes network diffusion principles and tools[107] to infer the molecular mechanisms and signalling pathways which mediate the effect of microbial proteins on host responses as measured by transcriptomic or proteomic read-outs. Flexibility is provided for users to feed in the desired datasets at any given step of the pipeline. Given the advent of new computational tools in inter-species interactions and pipeline management platforms, it is expected that an increasing number of dedicated bioinformatic workflows for microbe-host interactions will be developed in the near future.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

5. Challenges

Over the past decade, various advances in the domain of computational analysis of microbe-host interactions have been made. However, despite this progress, there remain many challenges as described below. These challenges also present many opportunities and the need to come up with innovative approaches and solutions.

5.1 Catching up with complex infection processes

Infection biology has taken new strides over the past years with new molecule classes[99,108–111] and cell-types[112] being discovered as having a role in the infection process. With that, novel interaction types between various molecular classes are also unearthed[113]. In some cases, computational methods have not caught up with molecular mechanisms. Hence, computational method developments are always a step behind the complexity associated with infection biology. This gap is all the more prevalent for commensal organisms in contrast to pathogens due to the constant and historically prevalent study bias.

5.2 Lack of experimental datasets

Non-model organisms and non-pathogenic organisms such as probiotics and commensals also suffer from the fact that there exists a considerable knowledge gap in terms of known/experimentally verified molecular interactions. This affects the performance of computational methods considerably due to the need for large sets of true positives for the assessment of predictive algorithms[114]. In addition, this also influences the coverage and accuracy of interolog approaches since they harness already existing true positive datasets for extrapolating to the species-pairs of interest based on orthology.

5.3 False-positives

As with any computational algorithm, microbe-host interaction prediction methods also face the curse of false positives. This issue could be exacerbated by the availability of relatively small true positive (truly interacting) and true negative (non-interacting sets) datasets[114]. Furthermore, the evolutionary distance and difference in infection process between the template species-pairs and the species-pair of interest in addition to the absence of orthologous molecular components involved in the interactions could also contribute to the inflated false positive rates, reduced performance and coverage.

5.4 Community-wide interaction prediction

Most of the microbe-host interaction computational tools with the exception of a handful of methods[105] including genome-scale metabolic modelling methods[25,26,27,28,29,30,115,116] have been directed at uncovering interactions corresponding to individual microbe-host pairs. This is a major drawback of existing methodologies, especially given the fact that phenotypes related to health and disease are associated with changes in community wide alterations in terms of microbial compositions[117–121].

5.5 Modelling dynamics of microbe-host interactions

Last but not the least, current methods involved in microbe-host interaction analysis are not equipped to handle the dynamic nature of natural ecosystems and ecological niches in which

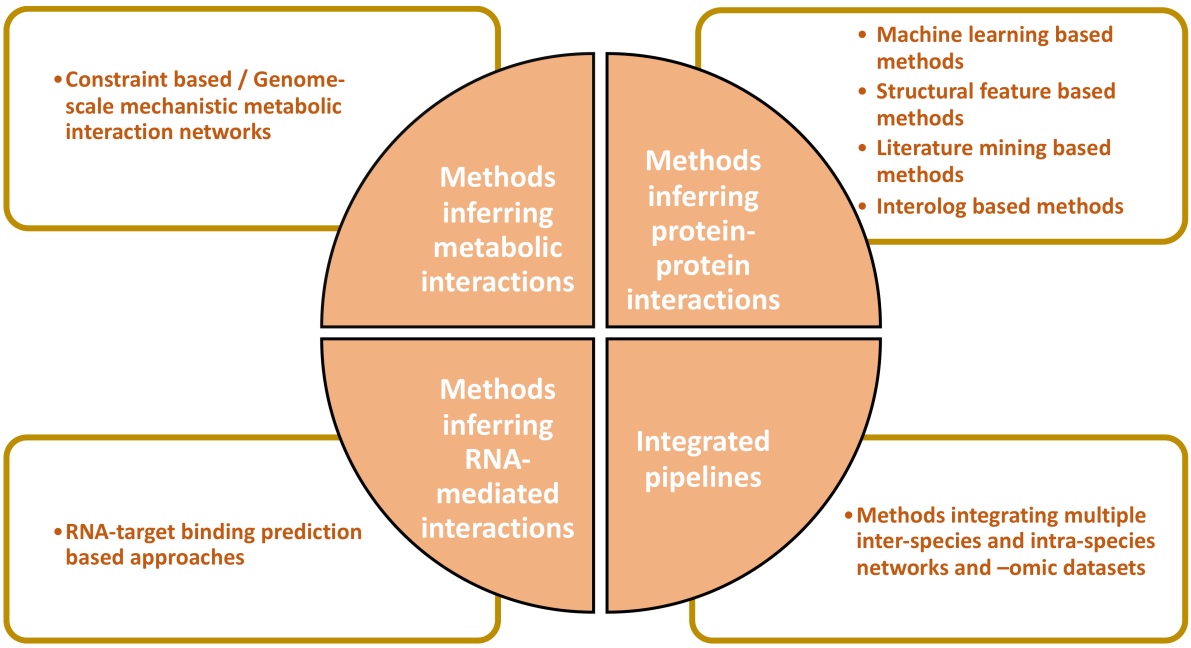
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

the interactions are embedded. Although it is a generic drawback of many bioinformatic approaches, given the need to accurately model microbe-host interactions, it is a challenge which needs coordinated efforts between modellers and experimental biologists, bioinformaticians.

6. Conclusion

Since the advent and expansion of high-throughput sequencing technologies, various observational studies of microbial communities inhabiting various ecological niches (inside host organisms for example) have been carried out. This has mostly resulted in associations with health- or disease-associated phenotypes. However, there is huge gap in terms of the mechanisms mediated by these microbial communities and how these mechanisms contribute to the observed phenotypes. Despite the availability of experimental datasets which capture some of these mechanisms such as PPIs, these are either confined to model organisms or well-studied pathogens. By enabling researchers to make de-novo inter-species molecular interactions and to extrapolate existing microbe-host interaction datasets to the species-pairs of interest, computational approaches provide researchers with the tools to upscale microbe-host interaction research. Although several limitations and caveats exist which need to be tackled on a case-by-case basis, computational methods tend to aid microbe-host interaction researchers by reducing the variable space, prioritizing interactions and eventually building hypothesis for further experimental verification.

Figure 1. Overview of the four different categories of computational methods which help infer the molecular mechanisms of microbe-host interactions.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

Table 1. Studies using genome-scale metabolic models and constraint based approaches to infer mechanistic co-metabolic interactions between microbial and host species.

Study	Context
- Rodenburg et al 2019[26]	Integrated metabolic model of *P. infestans* infecting tomato (*S. lycopersicum*)
- Islam et al 2019[25]	Genome-scale metabolic model between key members in the rumen microbiome and the viral phages
- Hertel et al 2019[115]	Integrated constraint-based model revealing microbe-host interactions in Parkinson's Disease
- Aller at et al 2018[116]	Genome-scale model integrating biochemical demands arising from virus production and human macrophage cell metabolism
- Ding et al 2016[28]	Simulation of co-metabolic model of different enteropathogens in response to various host environments
- Heinken and Thiele 2015[30]	In-silico microbe-host gut co-metabolic model to predict effects of different host dietary schemes
- Heinken et al 2013[29]	Experimentally validated gut co-metabolic model between commensal bacterium *B. thetaiotaomicron* and mouse
- Bordbar et al 2010[27]	*M. tuberculosis* infecting human alveolar macrophage supported by high-throughput data from infected conditions

Table 2. Computational approaches and methods inferring protein-protein interactions mediating inter-kingdom cross-talk between microbial and host organisms. *DDI – Domain-domain interaction; DMI – Domain-motif interaction; PSS – Pairwise structural similarity.

Method and corresponding studies	Reported use-case (host-microbe)
Machine learning based methods	
- Leite et al 2018[67]	Bacteria–phage
- Dyer et al 2011[122], Hongjaisee et al 2019[123],Shoombuatong et al 2012[63], Tastan et al 2009[124], Qi et al 2010[71], Nouretdinov et al 2012[125], Mei 2013[68]	Human–HIV
- Kshirsagar et al 2013[126]	Human–*F. tularensis,* Human–*Y. pestis,*
- Wuchty 2011[127]	Human–*B. anthracis,* Human-*S. typhi*
- Kosesoy et al 2019[66]	Human–*P. falciparum*
- Emamjomeh et al 2014[69], Kim et al 2017[64],Cui et al 2012[65]	Human–*Y. pestis,* Human–*B. anthracis*
- HOPITOR[128]	Human–Hepatitis C virus
- Liao et al 2011[129]	Generic (Human–virus PPIs)
- Mei et al 2018[62], Sun et al 2018[130]	Human–*S. japonicum*
- Kargarfard et al 2016[131]	Human–*M. tuberculosis*
- Kim et al 2017[64],Cui et al 2012[65],Dong et al 2015[132]	3 hosts and 674 influenza strains Human–Human papillomavirus
- Lai et al 2012[133]	
- Mei and Zhu 2014a[70]	Human–Influenza A virus
- Mei and Zhu 2014b[134]	Human–HTLV retroviruses

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

- Lian et al 2019[135]	Human–*Salmonella* Human–*Y. pestis*
Structural feature based methods (features used)	
- Dyer at al 2007[33] (DDI)	Human–*P. falciparum*
- Nourani et al 2016[136](DDI)	Human–multiple viruses
- Sudhakar et al 2019[45](DDI, DMI)	Human–multiple bacterial pathogens
- Dolittle and Gomez[54](PSS)	Human–Dengue virus, *A. aegyptii*–Dengue virus
- Cui et al 2016[56](PSS)	Human–HIV, Human–*M. tuberculosis*
- P-HIPSTer[58](PSS)	Human–multiple viruses
- Chen at al 2019[55](PSS)	Human–Dengue virus 2, Human–West Nile virus
- Guven-Maiorov et al 2017[53](Mimicry)	Human–*H. pylori*
- Mahajan and Mande 2017[34](DDI)	Human–*M. tuberculosis*
- Zhang et al 2017[44] (DMI)	Grass carp–Grass carp reovirus
- Mehrotra et al 2017[36] (PSS, DDI, localization)	Human–*L. interrogans*, Human–*L. biflexa*
- Yadav et al 2014[137](Computational docking)	Human–*B. malayi*
- Halehalli and Nagarajaram 2015[46] (DDI, DMI)	Human–multiple viruses
- Davis et al 2007[138](PSS)	Human–multiple pathogens
- SugarBindDB[139](glycan mediated PPIs)	Generic
- Rajasekharan et al 2013[57](PSS)	Human–Chandipura virus
- Carducci et al 2010[37](DDI)	Human–papillomavirus type 16
- Franzosa and Xia 2011[59](PSS, sequence identity)	Human–multiple viruses
- Mary et al 2016[140](Motif analysis)	Human–Dengue virus
- Sahu et al 2014[38](DDI)	Arabidopsis–*P.syringae*
- Zhou et al 2013[35](DDI)	Human–*M. tuberculosis*
- Dar et al 2017[141](PTM)	Human–Zika virus
- Kim et al 2008[39](DDI)	Rice–*X. oryzae*
- Kerr et al 2015[142](Computational docking)	Human-New world arenaviruses
- Evans et al 2009[47](DMI)	Human–HIV
- Doxey and McConkey 2013[51] (Mimicry)	Human–multiple bacterial pathogens
- Mei and Zhang et al 2020[52] (Mimicry)	Human–*S. typhimurium* and *Human respiratory syncytial virus*
Data/Literature mining based methods	
- Thieu et al 2012[77]	Generic
- Viruses.STRING[76]	319 hosts and 239 viruses
- Li et al 2018[78]	Human–Epstein-Barr virus
- Saik et al 2016[79]	Human– Hepatitis C virus
- Garcia-Perez et al 2018[80]	Human–Influenza A virus
- Mondal et al 2012[81], Mukhopadhyay et al 2012[82], Ray et al 2012[83]	Human–HIV
"Interolog" based methods	
- Krishnadev and Srinivasan 2008[85], Lee et al 2008[86]	Human–*P. falciparum*
- Krishnadev and Srinivasan 2011[87]	Human–*E. coli,* Human– *S. typhimurium,* Human–*Y. pestis*
- Tyagi et al 2009[89]	Human–*H. pylori*
- Cui et al 2016[56]	Human–HIV, Human–*M. tuberculosis*
- Schleker et al 2012[88]	Human–*Salmonella, Salmonella–A. thaliana*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

- Li et al 2012[143] - Wallqvist et al 2017[91] - Cuesta-Astroz et al 2019[93] - Zhou et al 2014[90] - Wang et al 2013[144] - Barh et al 2013[92]	*A. thaliana–R. solanacearum* Human–*C. burnetti* Human and 15 eukaryotic parasites Human–*M. tuberculosis* Zebrafish-*C. albicans* Human–*C. pseudotuberculosis,* Human–*C. diphtheriae,* Human–*M. tuberculosis,* Human–*C. ulcerans,* Human–*Y. pestis,* and Human–*E. coli*

Table 3. Examples of studies utilizing computational approaches to infer RNA-mediated interactions between microbes and hosts.

Study	Context
- Demirci and Adan 2020[101]	Analysis revealing the potential interactions between mature micro-RNA like viral RNA sequences and host genes
- ViRBase[94]	Source of experimentally verified virus–host non-coding RNA-associated interactions; also contains predicted binding sites of virus non-coding RNAs on host proteins and RNAs

Table 4. Integrated pipelines used to infer microbe-host interactions by combining heterogeneous -omic datasets.

Methodology	Functionalities
- MicrobioLink[105]	Integrating microbe-host protein interaction networks with host responses and host regulatory/signalling networks using network diffusion principles
- KBase[104]	Integrated platform enabling data sharing, integration, and analysis of -omic datasets from microbes, plants, and their communities by creating computational workflows
- Li et al 2015[145]	Identifying critical effectors involved in host-pathogen interactions by integrating multiple lines of -omic evidence

Bibliography

1. Braga RM, Dourado MN, Araújo WL. Microbial interactions: ecology in a molecular perspective. *Braz. J. Microbiol.* 2016;47 Suppl 1:86–98.
2. Martin AM, Yabut JM, Choo JM, et al. The gut microbiome regulates host glucose homeostasis via peripheral serotonin. *Proc. Natl. Acad. Sci. USA*. 2019;116:19802–19804.
3. Mendes V, Galvão I, Vieira AT. Mechanisms by which the gut microbiota influences cytokine production and modulates host inflammatory responses. *J. Interferon*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

Cytokine Res. 2019;39:393–409.
4. Pickard JM, Zeng MY, Caruso R, et al. Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. *Immunol. Rev.* 2017;279:70–89.
5. Ohland CL, Jobin C. Microbial activities and intestinal homeostasis: A delicate balance between health and disease. *Cell. Mol. Gastroenterol. Hepatol.* 2015;1:28–40.
6. Penny HA, Hodge SH, Hepworth MR. Orchestration of intestinal homeostasis and tolerance by group 3 innate lymphoid cells. *Semin Immunopathol*. 2018;40:357–370.
7. Hughes DT, Sperandio V. Inter-kingdom signalling: communication between bacteria and their hosts. *Nat. Rev. Microbiol.* 2008;6:111–120.
8. Levy SE, Myers RM. Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics. Hum. Genet.* 2016;17:95–115.
9. Valli RXE, Lyng M, Kirkpatrick CL. There is no hiding if you Seq: recent breakthroughs in Pseudomonas aeruginosa research revealed by genomic and transcriptomic next-generation sequencing. *J. Med. Microbiol.* 2020;69:162–175.
10. Jacob JJ, Veeraraghavan B, Vasudevan K. Metagenomic next-generation sequencing in clinical microbiology. *Indian J Med Microbiol*. 2019;37:133–140.
11. Shah P, Fritz JV, Glaab E, et al. A microfluidics-based in vitro model of the gastrointestinal human-microbe interface. *Nat. Commun.* 2016;7:11535.
12. May S, Evans S, Parry L. Organoids, organs-on-chips and other systems, and microbiota. *Emerg. Top. Life Sci.* 2017;1:385–400.
13. Eain MMG, Baginska J, Greenhalgh K, et al. Engineering Solutions for Representative Models of the Gastrointestinal Human-Microbe Interface. *Engineering*. 2017;3:60–65.
14. Roume H, Heintz-Buschart A, Muller EEL, et al. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *npj Biofilms and Microbiomes*. 2015;1:15007.
15. Muller EEL, Glaab E, May P, et al. Condensing the omics fog of microbial communities. *Trends Microbiol.* 2013;21:325–333.
16. Fritz JV, Desai MS, Shah P, et al. From meta-omics to causality: experimental models for human microbiome research. *Microbiome*. 2013;1:14.
17. Dix A, Vlaic S, Guthke R, et al. Use of systems biology to decipher host-pathogen interaction networks and predict biomarkers. *Clin. Microbiol. Infect.* 2016;22:600–606.
18. Martinez KB, Leone V, Chang EB. Microbial metabolites in health and disease: Navigating the unknown in search of function. *J. Biol. Chem.* 2017;292:8553–8559.
19. Wong ACN, Vanhove AS, Watnick PI. The interplay between intestinal bacteria and host metabolism in health and disease: lessons from Drosophila melanogaster. *Dis. Model. Mech.* 2016;9:271–281.
20. Samal SS, Radulescu O, Weber A, et al. Linking metabolic network features to phenotypes using sparse group lasso. *Bioinformatics*. 2017;33:3445–3453.
21. Pey J, Tobalina L, Cisneros JPJ de, et al. A network-based approach for predicting key enzymes explaining metabolite abundance alterations in a disease phenotype. *BMC Syst. Biol.* 2013;7:62.
22. Zampieri G, Vijayakumar S, Yaneske E, et al. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* 2019;15:e1007084.
23. Mendoza SN, Olivier BG, Molenaar D, et al. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.* 2019;20:158.
24. Cho JS, Gu C, Han TH, et al. Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring. *Current Opinion*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

in Systems Biology. 2019;15:1–11.
25. Islam MM, Fernando SC, Saha R. Metabolic modeling elucidates the transactions in the rumen microbiome and the shifts upon virome interactions. *Front. Microbiol.* 2019;10:2412.
26. Rodenburg SYA, Seidl MF, Judelson HS, et al. Metabolic Model of the Phytophthora infestans-Tomato Interaction Reveals Metabolic Switches during Host Colonization. *MBio*. 2019;10.
27. Bordbar A, Lewis NE, Schellenberger J, et al. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol. Syst. Biol.* 2010;6:422.
28. Ding T, Case KA, Omolo MA, et al. Predicting Essential Metabolic Genome Content of Niche-Specific Enterobacterial Human Pathogens during Simulation of Host Environments. *PLoS One*. 2016;11:e0149423.
29. Heinken A, Sahoo S, Fleming RMT, et al. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes*. 2013;4:28–40.
30. Heinken A, Thiele I. Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes*. 2015;6:120–130.
31. Schweppe DK, Harding C, Chavez JD, et al. Host-Microbe Protein Interactions during Bacterial Infection. *Chem. Biol.* 2015;22:1521–1530.
32. Ding Z, Kihara D. Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features. *Curr Protoc Protein Sci*. 2018;93:e62.
33. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics*. 2007;23:i159-66.
34. Mahajan G, Mande SC. Using structural knowledge in the protein data bank to inform the search for potential host-microbe protein interactions in sequence space: application to Mycobacterium tuberculosis. *BMC Bioinformatics*. 2017;18:201.
35. Zhou H, Rezaei J, Hugo W, et al. Stringent DDI-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. *BMC Syst. Biol.* 2013;7 Suppl 6:S6.
36. Mehrotra P, Ramakrishnan G, Dhandapani G, et al. Comparison of Leptospira interrogans and Leptospira biflexa genomes: analysis of potential leptospiral-host interactions. *Mol. Biosyst.* 2017;13:883–891.
37. Carducci M, Licata L, Peluso D, et al. Enriching the viral-host interactomes with interactions mediated by SH3 domains. *Amino Acids*. 2010;38:1541–1547.
38. Sahu SS, Weirick T, Kaundal R. Predicting genome-scale Arabidopsis-Pseudomonas syringae interactome using domain and interolog-based approaches. *BMC Bioinformatics*. 2014;15 Suppl 11:S13.
39. Kim J-G, Park D, Kim B-C, et al. Predicting the interactome of Xanthomonas oryzae pathovar oryzae for target selection and DB service. *BMC Bioinformatics*. 2008;9:41.
40. Akiva E, Friedlander G, Itzhaki Z, et al. A dynamic view of domain-motif interactions. *PLoS Comput. Biol.* 2012;8:e1002341.
41. Gibson TJ, Dinkel H, Van Roey K, et al. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun. Signal.* 2015;13:42.
42. Idrees S, Pérez-Bercoff Å, Edwards RJ. SLiM-Enrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions. *PeerJ.* 2018;6:e5858.
43. Perkins JR, Diboun I, Dessailly BH, et al. Transient protein-protein interactions: structural, functional, and network properties. *Structure*. 2010;18:1233–1243.
44. Zhang A, He L, Wang Y. Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions. *BMC Bioinformatics*. 2017;18:145.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

45. Sudhakar P, Jacomin A-C, Hautefort I, et al. Targeted interplay between bacterial pathogens and host autophagy. *Autophagy*. 2019;15:1620–1633.
46. Halehalli RR, Nagarajaram HA. Molecular principles of human virus protein-protein interactions. *Bioinformatics*. 2015;31:1025–1033.
47. Evans P, Dampier W, Ungar L, et al. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics*. 2009;2:27.
48. Via A, Uyar B, Brun C, et al. How pathogens use linear motifs to perturb host cell networks. *Trends Biochem. Sci.* 2015;40:36–48.
49. Hurford A, Day T. Immune evasion and the evolution of molecular mimicry in parasites. *Evolution*. 2013;67:2889–2904.
50. Cusick MF, Libbey JE, Fujinami RS. Molecular mimicry as a mechanism of autoimmune disease. *Clin Rev Allergy Immunol*. 2012;42:102–111.
51. Doxey AC, McConkey BJ. Prediction of molecular mimicry candidates in human pathogenic bacteria. *Virulence*. 2013;4:453–466.
52. Mei S, Zhang K. In silico unravelling pathogen-host signaling cross-talks via pathogen mimicry and human protein-protein interaction networks. *Comput Struct Biotechnol J*. 2020;18:100–113.
53. Guven-Maiorov E, Tsai C-J, Ma B, et al. Prediction of Host-Pathogen Interactions for Helicobacter pylori by Interface Mimicry and Implications to Gastric Cancer. *J. Mol. Biol.* 2017;429:3925–3941.
54. Doolittle JM, Gomez SM. Mapping protein interactions between Dengue virus and its human and insect hosts. *PLoS Negl. Trop. Dis.* 2011;5:e954.
55. Chen J, Sun J, Liu X, et al. Structure-based prediction of West Nile virus-human protein-protein interactions. *J Biomol Struct Dyn*. 2019;37:2310–2321.
56. Cui T, Li W, Liu L, et al. Uncovering New Pathogen-Host Protein-Protein Interactions by Pairwise Structure Similarity. *PLoS One*. 2016;11:e0147612.
57. Rajasekharan S, Rana J, Gulati S, et al. Predicting the host protein interactors of Chandipura virus using a structural similarity-based approach. *Pathog Dis*. 2013;69:29–35.
58. Lasso G, Mayer SV, Winkelmann ER, et al. A Structure-Informed Atlas of Human-Virus Interactions. *Cell*. 2019;178:1526–1541.e16.
59. Franzosa EA, Xia Y. Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. USA*. 2011;108:10538–10543.
60. Shastry KA, Sanjay HA. Machine learning for bioinformatics. In: Srinivasa KG, Siddesh GM, Manisekhar SR, eds. *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*. Algorithms for intelligent systems. Singapore: Springer Singapore; 2020:25–39.
61. Zhang M, Su Q, Lu Y, et al. Application of Machine Learning Approaches for Protein-protein Interactions Prediction. *Med Chem*. 2017;13:506–514.
62. Mei S, Flemington EK, Zhang K. Transferring knowledge of bacterial protein interaction networks to predict pathogen targeted human genes and immune signaling pathways: a case study on M. tuberculosis. *BMC Genomics*. 2018;19:505.
63. Shoombuatong W, Hongjaisee S, Barin F, et al. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput Biol Med*. 2012;42:885–889.
64. Kim B, Alguwaizani S, Zhou X, et al. An improved method for predicting interactions between virus and human proteins. *J Bioinform Comput Biol*. 2017;15:1650024.
65. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*. 2012;13 Suppl 7:S5.
66. Kösesoy İ, Gök M, Öz C. A new sequence based encoding for prediction of host-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

pathogen protein interactions. *Comput Biol Chem*. 2019;78:170–177.
67. Leite DMC, Brochet X, Resch G, et al. Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics*. 2018;19:420.
68. Mei S. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS One*. 2013;8:e79606.
69. Emamjomeh A, Goliaei B, Zahiri J, et al. Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol. Biosyst.* 2014;10:3147–3154.
70. Mei S, Zhu H. Computational reconstruction of proteome-wide protein interaction networks between HTLV retroviruses and Homo sapiens. *BMC Bioinformatics*. 2014;15:245.
71. Qi Y, Tastan O, Carbonell JG, et al. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics*. 2010;26:i645-52.
72. Kumar R, Nanduri B. HPIDB--a unified resource for host-pathogen interactions. *BMC Bioinformatics*. 2010;11 Suppl 6:S16.
73. Singh N, Bhatia V, Singh S, et al. MorCVD: A Unified Database for Host-Pathogen Protein-Protein Interactions of Cardiovascular Diseases Related to Microbes. *Sci. Rep.* 2019;9:4039.
74. Gao NL, Zhang C, Zhang Z, et al. MVP: a microbe-phage interaction database. *Nucleic Acids Res.* 2018;46:D700–D707.
75. Durmuş Tekir S, Çakır T, Ardiç E, et al. PHISTO: pathogen-host interaction search tool. *Bioinformatics*. 2013;29:1357–1358.
76. Cook HV, Doncheva NT, Szklarczyk D, et al. Viruses.STRING: A Virus-Host Protein-Protein Interaction Database. *Viruses*. 2018;10.
77. Thieu T, Joshi S, Warren S, et al. Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics*. 2012;28:867–875.
78. Li C-W, Jheng B-R, Chen B-S. Investigating genetic-and-epigenetic networks, and the cellular mechanisms occurring in Epstein-Barr virus-infected human B lymphocytes via big data mining and genome-wide two-sided NGS data identification. *PLoS One.* 2018;13:e0202537.
79. Saik OV, Ivanisenko TV, Demenkov PS, et al. Interactome of the hepatitis C virus: Literature mining with ANDSystem. *Virus Res.* 2016;218:40–48.
80. García-Pérez CA, Guo X, Navarro JG, et al. Proteome-wide analysis of human motif-domain interactions mapped on influenza a virus. *BMC Bioinformatics*. 2018;19:238.
81. Anon. Prediction of protein interactions on hiv-1–human ppi data using a novel closure-based integrated approach. In: *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*. SciTePress - Science and and Technology Publications; 2012:164–173.
82. Mukhopadhyay A, Maulik U, Bandyopadhyay S. A novel biclustering approach to association rule mining for predicting HIV-1-human protein interactions. *PLoS One.* 2012;7:e32289.
83. Ray S, Mukhopadhyay A, Maulik U. Predicting annotated HIV-1-Human PPIs using a biclustering approach to association rule mining. In: *2012 Third International Conference on Emerging Applications of Information Technology*. IEEE; 2012:28–31.
84. Kshirsagar M, Schleker S, Carbonell J, et al. Techniques for transferring host-pathogen protein interactions knowledge to new tasks. *Front. Microbiol.* 2015;6:36.
85. Krishnadev O, Srinivasan N. A data integration approach to predict host-pathogen

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

protein-protein interactions: application to recognize protein interactions between human and a malarial parasite. *In Silico Biol (Gedrukt)*. 2008;8:235–250.
86. Lee S-A, Chan C, Tsai C-H, et al. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*. 2008;9 Suppl 12:S11.
87. Krishnadev O, Srinivasan N. Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int. J. Biol. Macromol.* 2011;48:613–619.
88. Schleker S, Garcia-Garcia J, Klein-Seetharaman J, et al. Prediction and comparison of Salmonella-human and Salmonella-Arabidopsis interactomes. *Chem Biodivers*. 2012;9:991–1018.
89. Tyagi N, Krishnadev O, Srinivasan N. Prediction of protein-protein interactions between Helicobacter pylori and a human host. *Mol. Biosyst.* 2009;5:1630–1635.
90. Zhou H, Gao S, Nguyen NN, et al. Stringent homology-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. *Biol. Direct*. 2014;9:5.
91. Wallqvist A, Wang H, Zavaljevski N, et al. Mechanisms of action of Coxiella burnetii effectors inferred from host-pathogen protein interactions. *PLoS One*. 2017;12:e0188071.
92. Barh D, Gupta K, Jain N, et al. Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli targeted by Piper betel compounds. *Integr Biol (Camb)*. 2013;5:495–509.
93. Cuesta-Astroz Y, Santos A, Oliveira G, et al. Analysis of Predicted Host-Parasite Interactomes Reveals Commonalities and Specificities Related to Parasitic Lifestyle and Tissues Tropism. *Front. Immunol.* 2019;10:212.
94. Li Y, Wang C, Miao Z, et al. ViRBase: a resource for virus-host ncRNA-associated interactions. *Nucleic Acids Res.* 2015;43:D578-82.
95. Agliano F, Rathinam VA, Medvedev AE, et al. Long Noncoding RNAs in Host-Pathogen Interactions. *Trends Immunol.* 2019;40:492–510.
96. Huang C-Y, Wang H, Hu P, et al. Small RNAs - Big Players in Plant-Microbe Interactions. *Cell Host Microbe*. 2019;26:173–182.
97. Ahmadi Badi S, Bruno SP, Moshiri A, et al. Small RNAs in Outer Membrane Vesicles and Their Function in Host-Microbe Interactions. *Front. Microbiol.* 2020;11:1209.
98. Weiberg A, Wang M, Bellinger M, et al. Small RNAs: a new paradigm in plant-microbe interactions. *Annu. Rev. Phytopathol.* 2014;52:495–516.
99. Duval M, Cossart P, Lebreton A. Mammalian microRNAs and long noncoding RNAs in the host-bacterial pathogen crosstalk. *Semin. Cell Dev. Biol.* 2017;65:11–19.
100. Shirahama S, Miki A, Kaburaki T, et al. Long Non-coding RNAs Involved in Pathogenic Infection. *Front. Genet.* 2020;11:454.
101. Saçar Demirci MD, Adan A. Computational analysis of microRNA-mediated interactions in SARS-CoV-2 infection. *PeerJ*. 2020;8:e9369.
102. Allmer J, Allmer J, Saçar Demirci MD. izMiR: computational ab initio microRNA detection. *Protoc exch*. 2016.
103. Dai X, Zhuang Z, Zhao PX. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* 2018;46:W49–W54.
104. Arkin AP, Cottingham RW, Henry CS, et al. Kbase: the united states department of energy systems biology knowledgebase. *Nat. Biotechnol.* 2018;36:566–569.
105. Andrighetti T, Bohar B, Lemke N, et al. MicrobioLink: An Integrated Computational Pipeline to Infer Functional Effects of Microbiome-Host Interactions. *Cells*. 2020;9.
106. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 2018;46:W329–W337.
107. Paull EO, Carlin DE, Niepel M, et al. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*. 2013;29:2757–2764.
108. Acuña SM, Floeter-Winter LM, Muxel SM. MicroRNAs: Biological Regulators in Pathogen-Host Interactions. *Cells*. 2020;9.
109. Katiyar-Agarwal S, Jin H. Role of small RNAs in host-microbe interactions. *Annu. Rev. Phytopathol.* 2010;48:225–246.
110. Rana A, Ahmed M, Rub A, et al. A tug-of-war between the host and the pathogen generates strategic hotspots for the development of novel therapeutic interventions against infectious diseases. *Virulence*. 2015;6:566–580.
111. Peters JM, Solomon SL, Itoh CY, et al. Uncovering complex molecular networks in host–pathogen interactions using systems biology. *Emerg. Top. Life Sci.* 2019;3:371–378.
112. Chattopadhyay PK, Roederer M, Bolton DL. A deadly dance: the choreography of host-pathogen interactions, as revealed by single-cell technologies. *Nat. Commun.* 2018;9:4638.
113. Silmon de Monerri NC, Kim K. Pathogens hijack the epigenome: a new twist on host-pathogen interactions. *Am. J. Pathol.* 2014;184:897–911.
114. Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* 2016;4:320–330.
115. Hertel J, Harms AC, Heinken A, et al. Integrated Analyses of Microbiome and Longitudinal Metabolome Data Reveal Microbial-Host Interactions on Sulfur Metabolism in Parkinson's Disease. *Cell Rep.* 2019;29:1767–1777.e8.
116. Aller S, Scott A, Sarkar-Tyson M, et al. Integrated human-virus metabolic stoichiometric modelling predicts host-based antiviral targets against Chikungunya, Dengue and Zika viruses. *J. R. Soc. Interface*. 2018;15.
117. Wang B, Yao M, Lv L, et al. The human microbiota in health and disease. *Engineering*. 2017;3:71–82.
118. Dominguez-Bello MG, Godoy-Vitorino F, Knight R, et al. Role of the microbiome in human development. *Gut*. 2019;68:1108–1114.
119. Bailey MA, Holscher HD. Microbiome-Mediated Effects of the Mediterranean Diet on Inflammation. *Adv. Nutr.* 2018;9:193–206.
120. Clemente JC, Ursell LK, Parfrey LW, et al. The impact of the gut microbiota on human health: an integrative view. *Cell*. 2012;148:1258–1270.
121. Koboziev I, Reinoso Webb C, Furr KL, et al. Role of the enteric microbiota in intestinal homeostasis and inflammation. *Free Radic. Biol. Med.* 2014;68:122–133.
122. Dyer MD, Murali TM, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect. Genet. Evol.* 2011;11:917–923.
123. Hongjaisee S, Nantasenamat C, Carraway TS, et al. HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. *Comput Biol Chem*. 2019;80:419–432.
124. Tastan O, Qi Y, Carbonell JG, et al. Prediction of interactions between HIV-1 and human proteins by information integration. *Pac Symp Biocomput*. 2009:516–527.
125. Nouretdinov I, Gammerman A, Qi Y, et al. Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. *Pac Symp Biocomput*. 2012:311–322.
126. Kshirsagar M, Carbonell J, Klein-Seetharaman J. Multitask learning for host-pathogen protein interactions. *Bioinformatics*. 2013;29:i217-26.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020

127. Wuchty S. Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens. *PLoS One*. 2011;6:e26960.
128. Basit AH, Abbasi WA, Asif A, et al. Training host-pathogen protein-protein interaction predictors. *J Bioinform Comput Biol*. 2018;16:1850014.
129. Liao Q, Yuan X, Xiao H, et al. Identifying Schistosoma japonicum excretory/secretory proteins and their interactions with host immune system. *PLoS One*. 2011;6:e23786.
130. Sun J, Yang L-L, Chen X, et al. Integrating Multifaceted Information to Predict Mycobacterium tuberculosis-Human Protein-Protein Interactions. *J. Proteome Res.* 2018;17:3810–3823.
131. Kargarfard F, Sami A, Mohammadi-Dehcheshmeh M, et al. Novel approach for identification of influenza virus host range and zoonotic transmissible sequences by determination of host-related associative positions in viral genome segments. *BMC Genomics*. 2016;17:925.
132. Dong Y, Kuang Q, Dai X, et al. Improving the Understanding of Pathogenesis of Human Papillomavirus 16 via Mapping Protein-Protein Interaction Network. *Biomed Res. Int.* 2015;2015:890381.
133. Lai Y-H, Li Z-C, Chen L-L, et al. Identification of potential host proteins for influenza A virus based on topological and biological characteristics by proteome-wide network approach. *J. Proteomics*. 2012;75:2500–2513.
134. Mei S, Zhu H. AdaBoost based multi-instance transfer learning for predicting proteome-wide interactions between Salmonella and human proteins. *PLoS One*. 2014;9:e110488.
135. Lian X, Yang S, Li H, et al. Machine-Learning-Based Predictor of Human-Bacteria Protein-Protein Interactions by Incorporating Comprehensive Host-Network Properties. *J. Proteome Res.* 2019;18:2195–2205.
136. Nourani E, Khunjush F, Durmuş S. Computational prediction of virus-human protein-protein interactions using embedding kernelized heterogeneous data. *Mol. Biosyst.* 2016;12:1976–1986.
137. Yadav S, Gupta S, Selvaraj C, et al. In silico and in vitro studies on the protein-protein interactions between Brugia malayi immunomodulatory protein calreticulin and human C1q. *PLoS One*. 2014;9:e106413.
138. Davis FP, Barkan DT, Eswar N, et al. Host pathogen protein interactions predicted by comparative modeling. *Protein Sci.* 2007;16:2585–2596.
139. Mariethoz J, Khatib K, Alocci D, et al. SugarBindDB, a resource of glycan-mediated host-pathogen interactions. *Nucleic Acids Res.* 2016;44:D1243-50.
140. Asnet Mary J, Paramasivan R, Shenbagarathai R. Identification of sequence motifs involved in Dengue virus-host interactions. *J Biomol Struct Dyn*. 2016;34:676–687.
141. Dar HA, Zaheer T, Paracha RZ, et al. Structural analysis and insight into Zika virus NS5 mediated interferon inhibition. *Infect. Genet. Evol.* 2017;51:143–152.
142. Kerr SA, Jackson EL, Lungu OI, et al. Computational and Functional Analysis of the Virus-Receptor Interface Reveals Host Range Trade-Offs in New World Arenaviruses. *J. Virol.* 2015;89:11643–11653.
143. Li Z-G, He F, Zhang Z, et al. Prediction of protein-protein interactions between Ralstonia solanacearum and Arabidopsis thaliana. *Amino Acids*. 2012;42:2363–2371.
144. Wang Y-C, Lin C, Chuang M-T, et al. Interspecies protein-protein interaction network construction for characterization of host-pathogen interactions: a Candida albicans-zebrafish interaction study. *BMC Syst. Biol.* 2013;7:79.
145. Li W, Fan X, Long Q, et al. Mycobacterium tuberculosis effectors involved in host-pathogen interaction revealed by a multiple scales integrative pipeline. *Infect. Genet. Evol.* 2015;32:1–11.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2020