In silico prediction of phage-bacteria infection networks as a tool to implement personalized phage therapy

Authors and affiliations

Yok-Ai Que, MD-PhD
Department of Intensive Care Medicine, Lausanne University Hospital Center (CHUV), CH-1011 Lausanne, Switzerland.

Prof. Carlos Peña, PhD
Institute for Information and Communication Technologies, University of Applied Sciences and Arts of Western Switzerland, HEIG-VD, CH-1401 Yverdon-Les-Bains, Switzerland.

Grégory Resch, PhD
Department of Fundamental Microbiology, University of Lausanne, CH-1015 Lausanne, Switzerland.

1. SUMMARY

The emergence and rapid dissemination of antibiotic resistance worldwide threatens medical progress. As a consequence, medicine might face a return to the pre-antibiotic era in a near future. The paucity of potential new anti-infectives in the pipeline of pharmaceutical industries urges the need for alternatives to fight this public health problem. Phage therapy might represent such an alternative. This re-emerging therapy uses viruses that specifically infect and kill bacteria during their life cycle to reduce/eliminate bacterial load and cure infections. These viruses, called bacteriophages or phages, have been co-evolving with bacteria for billions of years, controlling bacterial populations and epidemics, and contributing to their genetic exchanges. With the advantage of having low impact on the commensal flora, as they are highly strain specific, some phages might, nevertheless, harbor virulence factors and drive horizontal gene transfer mediating dissemination of pathogenic traits including antibiotic resistance, calling for careful selection before their therapeutic use (see below "*Phage lifestyle inside the bacteria*")**.**

The success of phage therapy mainly relies on the exact matching between both the target pathogenic bacteria and the therapeutic phage. Therefore, having access to a fully-characterized phage library is necessary, although not sufficient, to start with phage therapy. An essential, and obligate, second step to conceive personalized phage therapy treatments is the capacity to predict the interactions between the target pathogen and its potential phage. The long term goal of the proposed research is, therefore, to develop quantitative and predictive *in silico* models of phage-bacteria infection networks. These models will describe the interactions between phage and bacteria and will serve to fasten the selection of effective phages to propose phage therapy in a personalized fashion.

To efficiently predict successful phage-bacteria interactions suitable for phage therapy, we will develop a novel *in silico* methodology that will, ultimately, enable the selection of phage candidates from an existing phage library to target a given pathogenic bacteria. To achieve this, we will combine genomic information with state-of-the-art bioinformatic and machine learning techniques, taking advantage of the growing amount of interaction data already available as well as of our own data to keep uncovering new phage families. We will ensure that our methodology brings explanatory power along, thereby shedding light on the relevant genomic features underscoring the interactions. To challenge our approach, we will, eventually, validate prospectively our methodology using paradigmatic pathogens (Pendleton et al. 2013). For this, we will construct the phage-bacterium infection networks around those pathogens, to identify single phages and/or phage cocktails with extended bactericidal activities, as assessed in different models of infections, including a *Galleria melonnella* model of infection and a rat endocarditis model. We expect our methodology to drive a paradigm shift in phage therapy, by offering a time-sparing and easy-to-use way to accurately select phages for each individual patient. The methodology will be made available online and several future developments of this project are already envisioned.

2. RESEARCH PLAN

"Two major ways that modern medicine saves lives are through antibiotic treatment of severe infections and the performance of medical and surgical procedures under the protection of antibiotics" (Nathan & Cars 2014). The rapid spread of antibiotic resistance threatens medical progress and drifts us into a post-antibiotic era, where common infections or minor injuries will kill. The number of new antimicrobials under clinical development remains extremely limited and their progression into the clinical market is disappointingly slow, highlighting the urgency for innovative approaches in the management of multidrug resistant infections.

In this context, phage therapy encounters a high renewed interest in the Western world. Phages are ubiquitous environmental bacterial viruses. Soon after their discovery in the beginning of the 20th century and because of their capacity to kill bacteria, they were used as therapeutic agents for the treatment of human bacterial infections. Starting from the middle of the 20th century, the overwhelming success of antibiotic agents shadowed phage therapy in Occident. However, several Eastern European countries continued to produce therapeutic phage preparations and millions of patients suffering from a great variety of infectious diseases have reportedly been, and still are, successfully treated principally in Georgia, Russia, and Poland with no serious side effects reported so far.

Phage therapy success relies on the correct matching of a bacteria and a phage selected among a fully characterized phage library. Currently, rules for phage selection are empirical, but progress has recently been made enabling the use of *in silico* algorithms.

2.1 Current state of research in the field

Phages are found in natural and man-made environments, rapidly co-evolving with their bacterial targets. Being highly specific, their infectivity can actually vary drastically across bacterial strains of a same species (Flores et al. 2011). Determining phages host ranges is currently achieved using infection tests (Weitz et al. 2013) which, depending on the size of the reference bacterial panel to be tested, may take several days of lab work. Many positive interactions have been uncovered using these tests, revealing a modular structure at large phylogenetic scales due to the specificity of the phage for a particular bacterial species. At smaller scales however, a nested structure prevails, where highly phage-sensitive bacteria get infected by phages with both narrow and broad host range, whereas highly phage-resistant bacteria are only infected by broad range phages (Beckett et al. 2013). Despite these efforts including ~300 phage and bacterial strains, there is still little understanding of which phages can actually infect which bacteria (Flores et al. 2011). Considering that >1500 phages and >40000 bacteria have already been sequenced[1] and that viruses and bacteria evolve rapidly (Sanjuan et al. 2010, Denamur & Matic 2006), we anticipate that an *in silico* approach would instead allow to efficiently predict phage-bacteria interactions and contribute to the construction of *dynamic* large-scale and well-characterized phage-bacteria infection networks.

[1] http://www.ncbi.nlm.nih.gov/genome

What could be the important factors determining the degree of infectivity of a phage against a specific bacterial strain? Several studies have investigated how phages actually infect bacteria (Labrie et al. 2010; Samson et al. 2013) and how bacteria defend against phage invasions (Seed 2015). The main mechanisms are reviewed hereafter.

- *How phages "enter" their target bacteria*. Phages have receptor-binding proteins (RBPs) that can recognize and bind specifically to receptors on the bacterial cell surface. These bacterial receptors have been experimentally identified in some cases and shown to generally involve both proteins and cell-wall glycopolymers (Rakhuba et al. 2010). Once attached to the surface of their bacterial host, phages inject their genomes inside the bacterial cytoplasm. Thus, only phage genomes actually enter target bacteria.
- *Phage lifestyle inside the bacteria***.** Phages can generally be classified into two categories (Fig. 1). On the one hand, virulent phages enter directly the "lytic cycle", which yields the formation of new phages and their release by cell membrane cleavage, which therefore kills the bacteria. On the other hand, temperate phages instead undergo the "lysogenic cycle", which integrates the injected genome into that of the host (called *prophage*), which therefore replicates at the pace of bacterial divisions, until conditions change and the phage eventually enters the lytic cycle (e.g. after cell damage). Recently, a machine-learning approach dubbed PHACTS (McNair et al. 2012) was successfully developed to automatically classify phages' lifestyle based on their protein sequences. The method is freely available online[2].

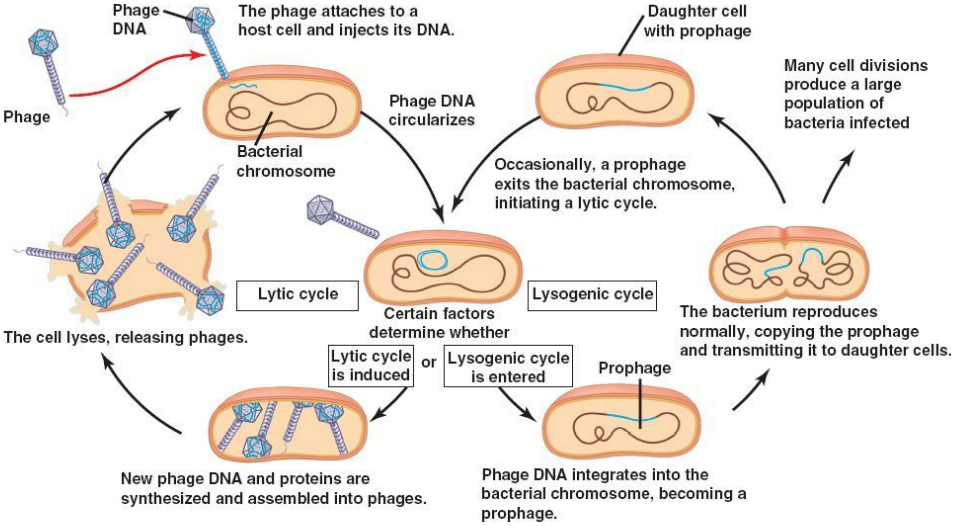
Phage DNA
The phage attaches to a host cell and injects its DNA.
Phage
Bacterial chromosome
Phage DNA circularizes
Daughter cell with prophage
Many cell divisions produce a large population of bacteria infected
Occasionally, a prophage exits the bacterial chromosome, initiating a lytic cycle.
Lytic cycle
Lysogenic cycle
The cell lyses, releasing phages.
Certain factors determine whether
Lytic cycle is induced
or
Lysogenic cycle is entered
Prophage
The bacterium reproduces normally, copying the prophage and transmitting it to daughter cells.
New phage DNA and proteins are synthesized and assembled into phages.
Phage DNA integrates into the bacterial chromosome, becoming a prophage.

Figure 1. Phage lifestyle (Reproduced from (Reece et al. 2010)).

- *How newly formed phages exit the bacteria*. During the lytic cycle, phage-encoded proteins get produced using the host machinery. Apart from the proteins that will eventually constitute the new virions, lysins also get produced. Lysins are the enzymes that will actually digest the bacteria's cell membrane by cleaving its surface peptidoglycans, enabling the release of the newly formed virions.

[2] http://www.phantome.org/PHACTS/

- *Bacterial defense mechanisms*. Bacteria have evolved sophisticated mechanisms to escape phage infection (Labrie et al. 2010) and phages have likewise developed strategies to defeat those bacterial defenses (Samson et al. 2013). In order to counteract phage adsorption, bacteria have for example evolved mutations in their phage receptors to render them unrecognizable to the phage. In other cases, phage receptors may be hidden behind physical barriers such as capsules (Seed 2015). Some bacteria also developed the ability to block phage DNA injection through superinfection exclusion systems, which block genome injection to the second phage trying to infect a bacterium. Other phages' potential weaknesses include genetically-encoded sites that could be targeted by the bacterial restriction-modification system which cuts foreign DNA at specific recognition sites. Another DNA cleavage system evolved by bacteria is the CRISPR-Cas system, which also involves recognition of a specific genomic region on the phage DNA (Garneau et al. 2010). Other bacteria choose to commit suicide in order to avoid phage dissemination into the clonal bacterial population (abortive infection system; Samson et al. 2013). Finally, phages can be defeated by bacteria through phage assembly interference, where bacteria encode phage-inducible chromosomal islands capable of negatively interacting with the assembly of the phage (Ram et al. 2012).

From the above, it appears that the specificity of a phage for a bacterial host is not only dictated by the phage receptor and lysins, but also greatly depends on the bacterial receptor and defense mechanisms.

Being able to predict phage-bacteria infection networks could be particularly beneficial in a clinical context for patients with multidrug resistant infections requiring rapid medical care. Indeed, phage therapy offers a promising alternative to the global crisis resulting from the emergence and dissemination of multiresistant bacterial strains. Current phage-based antimicrobial strategies include the use of a cocktail of lytic phages to specifically kill a population of bacteria (Matsuzaki et al. 2005), the use of purified lysins to lyse bacteria from the outside (Fischetti 2008), and the use of temperate phages to add extra genes to bacteria and render them antibiotic-sensitive (Edgar et al. 2012) (see (Yosef et al. 2014) for a review). With successes in preclinical and veterinary trials (Lu & Koeris 2011), phage therapy still misses access to well-characterized phage libraries as well as methods to rapidly screen for phage candidates given a bacterial strain.

Considering the advances in the understanding of phage biology and phage therapy, the increasing amount of genomic and interaction data available, the great successes of bioinformatics and machine learning to predict complex biological behaviors (Larrañaga et al. 2006), as well as the growing view that routine next-generation-sequencing will transform infectious disease management (Pak & Kasarkis 2015), it appears that using machine learning and bioinformatic tools to predict *in silico* phage-bacteria interactions based on genomic information is timely and will greatly accelerate further developments in the field by enabling the construction of comprehensive and dynamic phage-bacteria infection networks.

2.2 Current state of your own research

Yok Ai QUE, MD-PhD *FMH board certified in both Intensive Care and Internal Medicine - CHUV*

The core of the research of Dr Que's group is to conceive, elaborate and/or test novel strategies to combat complex infections, with the ultimate goal of uncovering novel strategies to fight the increasing threat of antibiotic resistance (Eggimann et al. 2015). Since a combination of innovative strategies, rather than a single approach, is required, the following complementary strategies are currently explored thanks to competitive grants:

- *From a Microbiome-related Approach to Synthetic Biology Treatments*

In the context of a multidisciplinary SystemsX.ch grant with the ETHZ, EPFL and UNIL, Dr Que is conducting a clinical trial to obtain a systems-level framework of microbial multispecies assemblies, their structuring, succession and functional behavior (e.g. gene expression) during burn wound sepsis. The outcome of this research, started in 2015, may lead to new perspectives and propositions of tests to possibly "guide" microbial community succession on damaged skin towards more healthy compositions exhibiting less risk of domination by extremely dangerous opportunistic pathogens (e.g., *Pseudomonas aeruginosa*).

- *Anti-Virulence Strategies: Interfering with Bacterial Quorum Sensing*

Quorum sensing (QS) is probably one of the best-studied examples of bacterial communication and *P. aeruginosa* is a paradigmatic microorganism to study QS. The applicant recently demonstrated that specialized *P. aeruginosa* QS molecules are heavily synthesized during human burn wound infections (Que et al. 2011) and that 2'-amino acetophenone, a QS molecule, mediates phenotypic changes in a sub-population of cells, which contribute to chronic infections, silence acute virulence functions of *P.* aeruginosa and the generation of antibiotic tolerant cells (Kesarwani et al. 2011; Bandyopadhaya et al. 2012; Que et al. 2013). Capitalizing on this previous work, Dr. Que participates in a large consortium, funded by SwissTransMed, that aims at designing new biological dressings that while improving wound healing, would also combat bacterial infections, in particular those caused by *P. aeruginosa*.

- *Antibiotherapy revisited: therapeutic dose monitoring (TDM)*

Profound modifications of metabolism and disturbed homeostasis observed among burn patients strongly complexify antibiotic prescription. Optimizing dosing levels, duration, route of administration, and use of combination drug therapy according to current PK/PD principles can suppress the emergence of resistance and minimize toxicity. Dr Que's group therefore conducted a retrospective study on the prescription of carbapenem in burn patients and demonstrated that their plasma levels are often insufficient resulting in treatment failures (Fournier et al. 2015). Through a two-year prospective, randomized, mono-centric, clinical trial, Dr. Que's group is currently analyzing the impact of systematic therapeutic drug monitoring (TDM) on anti-infective agent prescription amongst burned patients.

- *Evaluation of biomarkers to prognosticate outcome of septic patients.*

With the observed rising incidence of sepsis, risk stratification of septic patients is essential (i) to identify patients who are more likely to benefit from tailored advanced management of sepsis; (ii) to guide decision-making in institutions with scant resources; and (iii) to improve the selection of patients before their inclusion in interventional studies, based not only on the severity of infection but also mostly on their predicted outcome. In this context, Dr Que investigated the value of old (such as CRP (Que et al. 2015b)) and new biomarkers (such as pancreatic stone protein (Que et al. 2012; Que et al. 2015)) or genetic signatures (Yan et al. 2015) to predict either the outcome in ICU septic patients or the susceptibility to infections in burn patients.

- *Phage therapy: the Comeback of Natural Bacterial Predators*

Dr Que's group is deeply involved as a main partner in the PHAGOBURN study which aims at evaluating phage therapy for the treatment of *Escherichia coli* or *P. aeruginosa* burn wound infections. Recruitment started in July 2015. While collaborating at identifying new phages (Khalifa et al. 2015), Dr. Que's group evaluated *in vitro* and *in vivo* the efficacy of the anti-*P. aeruginosa* PHAGOBURN phage cocktail in models of endovascular infections (fibrin clots and experimental endocarditis in rats (Que et al. 2005; Moreillon and Que 2004)). Dr Que's research on phage therapy is financed by unrestricted grants from the Swiss National Foundation for education on phage therapy, the Loterie Romande to create a Phage Bank, and the FP7 European Commission program for the PHAGOBURN study.

Grégory Resch, PhD *Lecturer and Group Leader, Dpt of Fundamental Microbiology, Uni. of Lausanne*

Dr. Gregory Resch is specialized in bacteriophage research and has recently been awarded co-recipient for the ongoing phage therapy clinical trial PHAGOBURN (see above) as well as principal investigator for a SCOPES project and co-investigator for an AGORA project on phage therapy both funded by the SNSF. The 2-year SCOPES project, started in may 2014, aims at developing scientific exchange between Switzerland and Eastern countries, an opportunity that Dr. Resch used to start collaborating with Dr. Kutateladze, Director of the George Eliava Institute for Bacteriophages, Virology and Microbiology, Tbilisi, Georgia. In the frame of this project Dr. Resch and Dr. Kutateladze teams are developing a new phage cocktail against *Acinetobacter baumannii*. The ongoing AGORA project aims at discussing with the general public about phages and their therapeutic potential through workshops, exhibitions in public places and a dedicated website (www.phageback.ch). Dr. Resch published several manuscripts in both phage and lysin therapy fields and recently deposited a patent through the University of Lausanne to protect a new phage lysin active on different streptococcal species (European patent application N°13176730.3). In the past, Dr. Resch was recipient of a three-year Marie Curie Outgoing International Fellowship to work on the development of a phage lysin as a new antibacterial agent for which another patent has been deposited through the Rockefeller University, New York, US (WO2013052643 A1). His group currently focuses on the development of a rational approach to phage therapy involving rational design of phage cocktails (metagenomics of phage cocktails commercialized in Eastern countries, preclinical studies of diverse formulations), detailed study of bacterial resistance to therapeutic phages and impact of therapeutic phages on wound microbiomes. Of particular relevance to the present proposal, he has extensive experience with animal models of infectious diseases induced by various bacterial pathogens and is an authorized person for supervising animal experimentation (RESAL Module 1 + LTK Module 2).

Prof. Carlos PEÑA, PhD *University of Applied Sciences and Arts of Western Switzerland (HEIG-VD)*

The research conducted by the group of Prof. Peña at HEIG-VD focuses on the development of computational-intelligence methodologies and their application to real-world problems, specially those related with life sciences and biomedical engineering involving data analysis and predictive modeling. At the core of their research interests, their advanced computational methods derived from Artificial Intelligence and Machine Learning have proven very effective to tackle real-world problems in domains as diverse as agriculture, electrophysiology, smart buildings, and computational biology.

In this latter domain, they have conducted several projects relevant to the scope of this proposal, related with disease modeling, diagnostic-signature extraction from multi-omics data, population-based drug-concentration modelling and control, and bioinformatic analyses to predict robust epitopes.

- Between 2009 and 2011, they participated to the FP7 European project “PharMEA: Multi-Electrode Array technology based platform for industrial pharmacology and toxicology drug screening” that involved quasi-real-time processing and model-based analysis of very large amounts of electrophysiological data.
- Another project that illustrates their experience on predictive modeling is “ISyPeM: Intelligent Integrated Systems for Personalized Medicine” running from 2010 to 2013 and funded by the Nano-Tera Initiative. In this project, the HEIG-VD teams conceived a tool for personalized (stratified) prediction of drug concentration in blood, allowing thus to adapt drug dosage to optimize disease treatment.
- They have also conducted several projects on disease characterization and diagnostic based on high-content transcriptomics data (i.e., gene expression measurements). Among these projects, the SNSF-funded project “nanoFUGE: FUzzy modelling Gene-Expression data from Nanostring technology” conducted between 2010 and 2014 in partnership with the University of Geneva and the University Hospitals of Geneva, where they used advanced machine learning methods to select and characterize a diagnostic biomarker pool enabling to accurately classify 18 sub-types of Leukemia.
- The project DiagnoSuite: Suite of data analysis and data management tools for discovering, developing, deploying, and exploiting biomarker-based diagnostic screening tests, funded by the CTI agency from 2013 to 2014 that closed a fruitful collaboration, active since 2010, with the diagnostic startup companies Diagnoplex SA and Novigenix SA.
- Currently, they are part of the European Eurostars-2 project “Fishguard: Improving fish viral diseases monitoring through the development of fast, cost-effective and in-field screening tests”, together with two European SMEs, Bioscientia and BIOTEM from Poland and France, respectively. The HEIG-VD contributes with bioinformatics and machine learning expertise to discover and characterise the biomarkers that will be embedded in the screening test.

2.3 Detailed research plan

Aims and rationale of the project

With the long-term vision that phage therapy could be proposed in addition to antibiotics to treat complex infections, the main goal of this project is to develop an *in silico* methodology to predict and better understand phage-bacteria infection networks. Using an original and novel approach, we will predict phage-bacteria interactions based on available genomic data originating both from existing databases and from our own research. In parallel, we will also enrich the amount of available data by collecting and sequencing new phages and bacteria. This will help challenge and refine our methodology to achieve high performance, and will also contribute to the construction of a large well-characterized phage library.

Ultimately, our methodology should enable the cost- and time-effective discovery of new phages targeting a bacteria of interest (and vice versa). To demonstrate this potential, we will use our methodology in a prospective study to *in silico* construct a large-scale phage-bacteria infection network, using as input four important human bacterial pathogens (*A. baumannii, E. coli, P. aeruginosa* and *Klebsiella pneumoniae*) and all the currently available phages. In the resulting network, every edge between the bacterium and a phage will be labeled with a likelihood score, as well as with explanatory features underlying the interaction. This information will be used by the biologists and clinicians in the project to select the most likely (combination of) phages to use *in vivo*. Predicted phages will be tested *in vivo* in the *Galleria mellonella* insect model of infection. Only the most efficient phage cocktail and monophage will be further tested *in vivo* in bacteremic mice infected with each bacterial pathogen to validate the method. Finally, in order to prove the clinical significance of the prediction in a more sophisticated model, efficacy of the “best” cocktail for *P. aeruginosa* and *K. pneumoniae* will be compared to the most efficient monophage and standard of care in a rat model of experimental endocarditis (inflammation of the inner layer of the heart).

In summary, the specific aims of this project consist of the following parts (summarized in Fig. 2):

A. **Collection & genome sequencing, assembly and annotation:**
   1. Bacteria.
   2. Phages.

B. **Retrospective *in silico* prediction of phage-bacteria infection networks:**
   1. Develop a novel machine-learning-based methodology to predict phage-bacteria interactions based on sequence data alone.
   2. For each predicted interaction, identify the most relevant features underlying it (explanatory power of the methodology).

C. **Prospective study on bacterial pathogens to discover new interactions:**
   1. Construct large-scale phage-bacteria infection networks.
   2. *In vivo* model of Galleria mellonella.
   3. *In vivo* validation of new phage-bacteria interactions using animal models of bacterial infections.

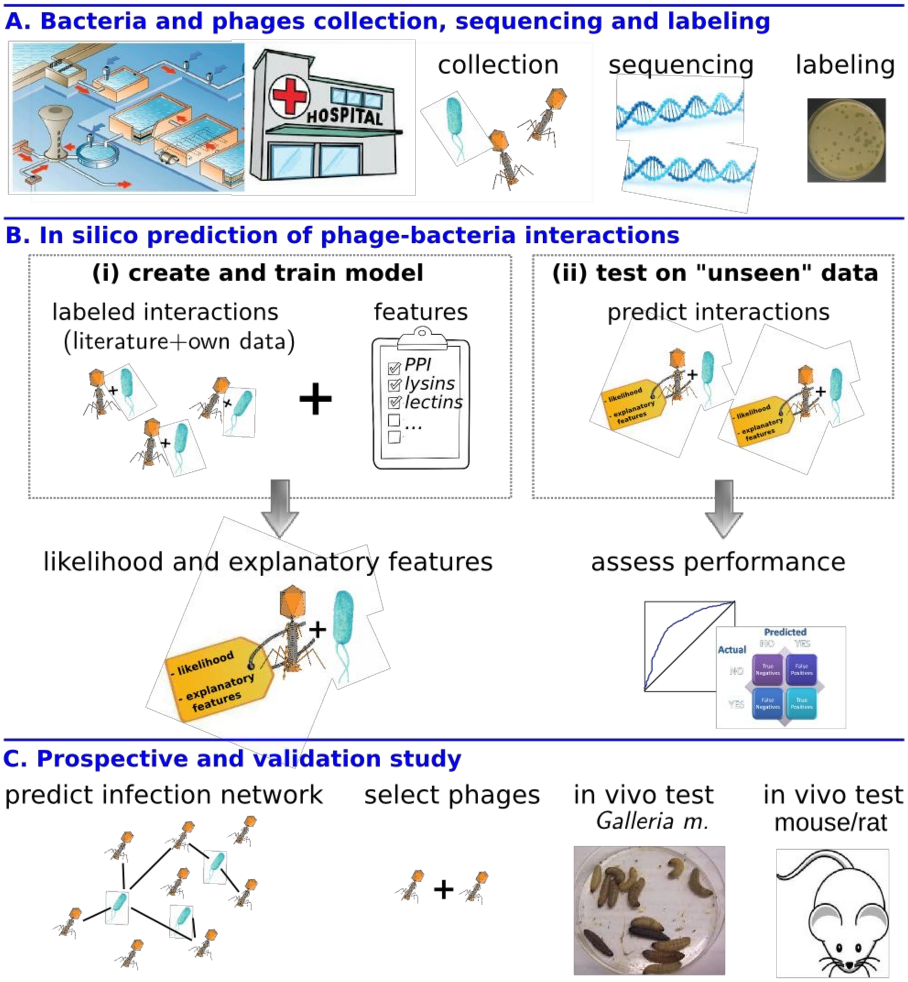

Figure 2: Aims of the project. Phages are represented in orange-brown and bacteria in blue (not drawn to scale). Refer to the text above for the legend.

A. Collection & genome sequencing, assembly and annotation

A.1. Bacteria

In order to get enough data to test and refine the algorithm, we will isolate phages specific to four important human bacterial pathogens, namely *A. baumannii, E. coli, P. aeruginosa* and *K. pneumoniae*.

Dr. Grégory Resch and Dr. Que will provide collections of >100 isolates/pathogen regrouping both clinical and laboratory strains. These isolates were collected previously and are currently stored at -80°C in glycerol stocks at the UNIL/CHUV. However, only a handful of strains were fully sequenced. Therefore, based on host ranges of isolated phages (see below), additional bacterial strains will be selected to be sequenced, assembled, and annotated. We plan to sequence at least 50 bacterial isolates for each pathogen, i.e. a total of 200 strains. Bacterial genomic DNAs will be prepared at UNIL/CHUV using commercialized kits. Library preparation (NexteraXT) and genome sequencing (MiSeq 150PE) will be performed at the Genomic Technologies Facility (GTF) of the Center for Integrative

Genomics (CIG). In order to significantly reduce sequencing costs, it is already planned to multiplex at least 25 samples in one lane. Assembly will be performed using the computing power of Vital-IT at UNIL. Dr. Resch has established strong connections through past collaborations with the GTF and has full access to Vital-IT. Annotation will be done automatically with help of the RAST Server (http://rast.nmpdr.org/) and nucleotide sequences will be deposited in public databases.

A.2. Phages

Our plan is to collect at least 20 lytic and 20 temperate phages for each pathogen, i.e. a total of 160 phages. Thanks to an unrestricted grant from the Loterie Romande, phages are currently routinely isolated in the laboratories of Drs. Resch and Que. To date, more than 15 unique lytic or temperate phages specific for *A. baumannii* and *P. aeruginosa* have been isolated, sequenced and annotated. Additional candidate phages active against *E. coli* and *K. pneumoniae* will be obtained (i) from wastewater samples of different wastewater plants in Switzerland and France (lytic and temperate phages), and (ii) by forcing prophages present in bacteria from our collections to switch to lytic mode and produce virions using mitomycin C induction (temperate phages). Potential phages in a sewage water sample will be first amplified overnight using growth medium containing the bacterial strains of interest. We will then use standard centrifugation and filtration to remove the bacteria and store the obtained suspension of phages at 4°C until further use (see e.g. Methods in (Peters et al. 2015)).

Antibacterial activity, revealing the presence of a phage specific for a given bacterium, will be screened through drop test. This technique consists in the deposition of an aliquot of the amplified sample on the surface of a soft agar layer containing the bacterium of interest. Appearance of clear zones after overnight incubation indicates presence of one or more specific phages (temperate or lytic). Individual phages are then further obtained using the classical double-layer assay consisting in mixing an aliquot of the amplified sample with the same bacterial strain on which a clear zone was observed in the drop test. After 24h incubation, clear plaques of different sizes form on the plates, from which phages can be isolated using standard methods. This process is repeated three times in order to obtain a single phage type within a single plaque. Each single phage is then further amplified in liquid culture to obtain batches containing from 10^8 to 10^{12} plaque forming units (PFU)/mL.

Phage genomic DNAs will be purified at UNIL using classical methods of phenol-chloroform precipitation. Of note, if we were to identify phages harboring RNA genomes (expected to be very rare), we would also include them through RNA sequencing. As for bacteria, library preparation (NexteraXT) and genome sequencing in multiplex (MiSeq 150PE) will be performed at the Genomic Technologies Facility (GTF) of the Center for Integrative Genomics (CIG). Assembly will be performed using the computing power of Vital-IT and annotation will be done automatically with help of the RAST Server (http://rast.nmpdr.org/). Nucleotide sequences will also be deposited in public databases.

B. Retrospective in silico prediction of phage-bacteria infection networks

B.1. Prediction of phage-bacteria interactions

Supervised machine learning consists in using features from labeled *training* examples to infer rules for mapping *new* examples. Classification of objects into two or more classes is a typical machine learning problem, where a "smart" combination of features, either existing in the original data or extracted from them, is identified by the model and used to discriminate between objects belonging to different classes (see Fig. 3). With this project, we aim at conceiving a machine learning method, or set of methods, able to build classifiers that predict whether or not a given pair of phage and bacterium can interact (and the likelihood of such a prediction).

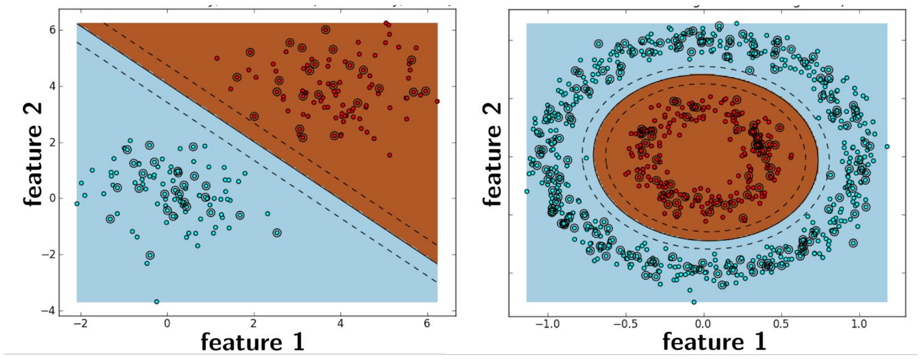

Figure 3. Cartoon example of a two-class problem with linearly-separable (left) and non-linearly-separable data. Image from http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html.

Datasets. We will take advantage of existing experimental knowledge to build our *training/validation* datasets. Indeed, Flores and co-workers (Flores et al. 2011) have already compiled a set of 37 studies with direct laboratory evidence of phage-bacteria interactions. Additionally, we will perform a systematic literature search to extend this dataset with new data. Importantly, since we are interested in predicting phage-bacteria interactions based on genomic data, we will also retrieve, whenever possible, the corresponding genomes of phages and bacteria from GenBank[3]. Moreover, annotated interaction data already available from the Resch lab will be added to the dataset. All this data will constitute the *positive* examples of our datasets. Equally important, we will include *negative* examples in our datasets, corresponding to known pairs of *non*-interacting phages and bacteria. In the case where the number of validated negative examples were to be too low (in comparison to the size of the positive dataset), we will randomly add pairs of phage-bacterium provided that they do not already appear in the positive dataset. This strategy has been used to generate negative datasets for predicting protein-protein interactions and is considered acceptable because the probability of committing an error while picking a random pair is low (Ben-Hur & Noble 2006).

[3] http://www.ncbi.nlm.nih.gov/genbank/

We will also build a *test* set with “first-seen” data consisting of newly collected bacterial strains and bacteriophages to assess the performance of our methodology in a statistically robust fashion. In particular, we will ensure that the test set does not contain bacterial strains or phages already present in the training/validation sets.

Feature selection. Given a list of features and a general mathematical model on how to combine them, a machine learning method may determine the importance of each feature for classifying correctly the objects from the training set. The selection of *potentially informative* features to be extracted from the examples is therefore a crucial step in any machine learning problem. Note that, generally speaking, the adequacy of a given set of features may be closely related with the machine learning method used to build the predicting classifiers, rendering even more important to count on a large set of candidate features. We will, thus, investigate several potential features, e.g.:

- *Enrichment in phage-bacteria protein-protein interactions (PPI).* The STRING database (db) (Snel et al. 2000) reports known and predicted intra-species PPI, currently covering >9Mio proteins from >2000 organisms[4]. Computational models to predict intra-species PPI have used features like protein sequence similarity, protein domain similarity, gene ontology, pathways and gene expression, recently being extended to between-species *host-pathogen* PPI as well (Dyer et al. 2007; Coelho et al. 2014; reviewed in Nourani et al. 2015). In this context, assuming that an interacting pair of phage-bacterium should be significantly enriched in interspecies PPI, we propose to use the informative features reported in (Coelho et al. 2014) to predict phages-bacteria interactions. These include protein sequence information (amino acids frequency), clusters of orthologous groups (STRING-db; Snel et al. 2000), gene ontology profiles (GO consortium; Harris et al. 2004) and enriched conserved domain pairs (DOMINE-db; Yellaboina et al. 2011).
- *Lysins subsequences.* Lysins are the enzymes that specifically cleave peptidoglycans at the bacterial surface, eventually killing the bacteria and releasing the newly produced virions. Because of their modular structure consisting of an evolutionarily-conserved catalytic domain and a bacterial specific domain, lysins can be easily identified in phage genomes using existing bioinformatic tools (see Methods in (Oliveira et al. 2013)). Several characteristics of their genomic sequence, like the presence of k-mer subsequences and the distribution of amino-acids, will be investigated and used as features to predict phage-bacteria interactions.
- *Comparison of lectins and polysaccharide binding proteins*. Phages infect bacteria by recognizing specific receptors at the bacterial surface, usually cell-wall glycopolymers (Samson et al. 2013). Phages therefore contain lectins, i.e. proteins capable of binding carbohydrate domains. Reciprocally, bacteria contain proteins involved in carbohydrate biosynthesis to build the glycopolymers. Using the PROSITE db of protein domains[5], we will identify potential lectins in the phage genomes as well as proteins involved in carbohydrate biosynthesis in the bacterial genome and investigate features they may share. For example, we will investigate their respective

[4] http://string-db.org/
[5] http://prosite.expasy.org/

distributions of amino acids and amino acids pairs, as well as the overlap of k-mer sequences. If promising, these features will be used to predict phage-bacteria interactions.

Models. We will implement various state-of-the-art modeling methods (e.g. artificial neural networks (ANN), deep learning, support vector machines (SVM), random forests, fuzzy logic) to build predictive models using the features discussed above (see (Hastie et al. 2009) for a review of models). In order to maximize model performance, we will then combine their predictions by applying ensemble-learning approaches (Polikar 2009), taking advantage of stacking to also get a final likelihood score of interaction. According to their knowledge representation, the models may be classified as *white-box, black-box,* or *grey-box.* In the present project we will use mainly models of the two latter classes, as explained below.

- *Black-box models:* This term refers to models that provide input-to-output predictive mapping but do not offer any insight on the patterns or mechanisms that allow for such prediction. ANNs and SVMs both belong to this family. ANNs are generally presented as systems of interconnected "neurons" (i.e., relatively simple transfer functions) which exchange messages between each other. The most common neuronal organization consists of three consecutive, interconnected layers (input, hidden, and output). The connections between neurons have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and, thus, capable of learning. The most recent advances in ANNs concern deep learning which is a novel approach based on a large number of neurons and layers. They allow for the construction of highly predictive models able to capture underlying patterns in the data being modelled while still offering excellent generalization. The same is true for SVMs, which are powerful predictors that can use complex kernels to non-linearly map the data onto a space where the different classes can actually be linearly separated. (Note that as deep-learning usually requires large datasets, its adequacy to this project will be evaluated during the modelling phase.)
- *Grey-box models:* This term refers to models in between the *black-box* and the *white-box* mechanistic models. They include e.g. decision trees and fuzzy logic, and have the advantage of combining high predictive power with higher explanatory power. Indeed, these methods can define ‘human-manageable” rules (fuzzy or crisp rules) to classify the data, which can then be used to better understand the features underscoring the predictions.

Performance. Robust cross-validation will be performed in order to prevent over-fitting and to optimize model selection. Moreover, in order to test the performance of our methodology, we will use an independent *test* dataset consisting of pairs of phages and bacteria collected, sequenced, and annotated in sections *A* and *B* of this proposal. Our method will thus be used to predict the likelihood of these phage-bacteria interactions. The final performance of our method will be thoroughly assessed using several metrics (sensitivity, specificity, accuracy, false discovery rate, F1-score, as well as the Matthews correlation coefficient to account for the likely case where the positive and negative classes were to be of very different sizes).

B.2. Identify explanatory features for each predicted pair of phage-bacterium

Machine learning is perhaps most known for its predictive power, but it can also be used to explain data. Indeed, just like any other model, it can be used both to gain more insight in the underlying processes, and to predict the behaviour of new data. Thus, after developing a robust and statistically meaningful predictive model of phage-bacteria interactions in section B.1, we will ensure that explanatory features can also be extracted in a "human-manageable" way, in order to better understand the most relevant features underscoring phage-bacteria interactions.

In section B.1, we will be using state-of-the-art modeling methods, some of which are inherently defined as to enable the understanding of the explanatory features (*grey-box models*). Some other models are however known to be less suited to explain data (*black-box models*), but may prove to be more powerful. Since predictive power is more of a quantitative trait than explanatory power, which is more qualitative, we will, in section B.1, select machine learning model(s) mainly aimed at maximizing predictive power. In the case where the resulting explanatory power was to be low, because of the final selection of models, we would then use some of the models exhibiting better explanatory power in order to allow biologists and clinicians to qualitatively assess the data and predictions, and extract useful information to advance knowledge in the field, even at the cost of a slightly reduced predictive power. This strategy will avoid trading off predictive power while ensuring that explanatory power is maintained, which is particularly important in a clinical context.

C. Prospective study on four ESKAPE species to discover new interactions

C.1. Construct large-scale phage-bacteria infection networks.

We will apply our methodology to construct comprehensive phage-bacteria infection networks using as input the genomes of four clinically relevant pathogens, namely *A. baumannii, E. coli, P. aeruginosa* and *K. pneumoniae,* and the corresponding phages isolated and sequenced in the lab. Phages available in public phage banks (Félix d'Hérelle Reference Center for Bacterial Viruses at University of Laval, Québec and Eliava Institute of Bacteriophages, Microbiology and Virology, Tbilisi, Georgia) will also be considered. This will be performed by testing *in silico* all the possible pairs of interaction between the bacteria and the phage genomes. The final predictions will be labeled with a likelihood score of interacting as well as with explanatory features underlying the interaction. This will allow biologists and clinicians to select at least six phage cocktails and two monophage preparations, in addition to two formulations predicted to be inefficient (negative control). Each cocktails will be validated *in vitro* using classical methods such as drop tests, plaque assays, time-kill and growth curves. The cocktails will be further tested and validated *in vivo*.

C.2. In vivo model of Galleria mellonella

The value of *Galleria mellonella* as model organism to study phage therapy efficacy has been recently highlighted in several studies in which standard protocols are detailed (Seed and Dennis 2009; Abbasifar et al. 2014; Olszak et al.

2015). This wax moth larvae model presents the advantage that it allows testing many experimental conditions or drugs in a cost-effective way and reduced amount of time compared to classical rodent models, thereby complying with the 3R's rule aiming at reducing animal use in scientific research. *Galleria mellonella* will therefore allow us to test many phage-bacteria interactions identified by our *in silico* prediction model that would otherwise require massive use of mice. We plan on testing at least 10 different conditions for each pathogen (i.e. 40 conditions in total, c.f. *C.1*). This will ensure that we quickly get an insight into the most relevant phage cocktails that would then be selected for further evaluation in mouse and/or rat models of infectious diseases.

Measurement of bacterial pathogen virulence and determination of the optimal inoculum will be performed as described in (Olszak et al. 2015). Both bacteria and bacteriophages will be administered to larvae by injection into the ventral side of the last pair of pseudopods. After injection, the larvae will be incubated at 37 °C and the effects of infection will be checked at 8, 24, 48, 72, and 96h post-injection by assessment of survival and macroscopic appearance. All experiments will be performed at least in triplicate (10 larvae per group). The experiments will be controlled by observation of uninfected larvae, sham-infected larvae, larvae receiving phage lysate only, and infected but phage untreated larvae. One phage cocktail for each pathogen leading to the highest rate of survival will be selected for further investigation in rodent models.

C.3. In vivo validation of new phage-bacteria interactions using animal models of bacterial infections.

1. Mouse model of bacteremia

In order to verify the accuracy of the *in silico* predictions, we will test the "best" predicted phage cocktail against one strain of each pathogen in a mouse model of bacteremia which allows rapid determination of efficacy of a treatment in a mammal background.

All protocols required to set up the *in vivo* model of bacteremia will be submitted for approval by the Committee on the Ethics of Animal Experiments of the Consumer and Veterinary Affairs, Department of the State of Vaud. Moreover, the mouse model will be carried out in strict accordance with the recommendations of the Swiss Federal Act on Animal Protection. A total of 92+72=164 mice will be used for this study (see below for detailed repartition).

a. Determination of LD_{90} and bacteremic status

In order to conduct the therapeutic experiment with the phage cocktail, we will first need to identify the optimal LD_{90} (lethal dose to kill 90% of the test population) for each strain. Briefly, the protocol consists in intraperitoneal injection (i.p.) of 100µl of the bacterial pathogen resuspended in suited buffer. Survival of mice will be followed for up to ten days and mice showing a weight loss of ≥15% of the initial weight will be systematically euthanized with CO_2. Using the formula for dichotomous variables (Dell et al. 2002), the animal sample size is estimated to be five per group. Since we expect to test at maximum three inoculums + NaCl 0.9% (control group) for each pathogen, we anticipate 5*4=20 animals per pathogen. Therefore a total of 80 animals will be required for the LD_{90} determination of the four pathogenic strains selected.

In order to verify the bacteremic status of mice (i.e. presence of bacteria in blood with spread in different organs) and since we expect that bacteremia will be reached in maximum 2h after bacterial challenge, the lungs, left side kidney and spleen will be removed aseptically from three additional animals for each pathogen at 2h post-infection, homogenized in saline solution and plated for colony counting as described in (Vouillamoz et al. 2013). For this experiment, 3*4=12 animals will be required, i.e. a grand total of 80+12=92 mice.

b. Evaluation of efficacy of the selected phage cocktails in bacteremic mice.

Mice will be infected with LD_{90} determined for each pathogen in 1.a (see above). Treatment or placebo will be injected in 50µl 2h after bacterial challenge through intravenous (i.v) injection. Survival of mice will be followed for up to ten days and mice showing a weight loss of ≥15% of the initial weight will be systematically euthanized with CO_2. The animal sample size is here estimated to be nine per group (formula in (Dell et al. 2002)), meaning that a total of 9*2*4=72 mice will be required.

2. *Rat experimental endocarditis model*

In order to challenge the efficacy of our interdisciplinary approach, the “best” cocktail for *P. aeruginosa* and *K. pneumoniae* will be compared to the most efficient monophage and standard of care in an *in vitro* and *in vivo* rat model of experimental endocarditis.

a. *In vitro* fibrin clot model

Plasma clots inoculated with bacteria will be prepared as previously described (Entenza et al. 2009) and suspended in a phage cocktail solution. Clots will be removed at 0, 6 and 24 h of incubation and prepared as described for colony counting (Entenza et al. 2009). The activity of the phage cocktail will be considered bactericidal whenever it kills ≥3 $\log_{10}$ CFU (colony forming units) of the starting inoculum in the clot.

b. *In vivo* model

All animal protocols will be reviewed and approved by the Cantonal Committee on Animal Experiments of the State of Vaud, Switzerland. A mixture of ketamine (75 mg/kg) and midazolam (5 mg/kg) anaesthetics will be administered to the animals before any surgical procedure.

The production of catheter-induced aortic vegetations and the installation of an i.v. line into the superior vena cava, connected to an infusion pump to deliver treatments, will be performed in female Wistar rats as described elsewhere (Héraïef & Freedman 1982, Fluckiger et al. 1994).

Throughout all experiments, body weight will be monitored and animals showing weight loss of ≥15% of the initial weight will be systematically euthanized with CO_2

b.1. Determination of ID_{90}

In order to conduct the therapeutic experiments, we will first need to identify the optimal ID_{90} (bacterial dose to infect 90% of the inoculated rats) for each strain. Twenty-four hours after catheterization, rats will be challenged i.v. with bacterial pathogens. Eighteen hours after challenge, rats will be euthanized. Infected vegetations will be collected, homogenized in saline buffer and plated for colony counting. We plan on testing three different inoculum sizes per pathogen and six rats/inoculum. Therefore, a total of 7*6*2=84 rats will be needed for these experiments.

b.2. Therapeutic experiments

Twenty-four hours after catheterization, rats will be challenged i.v. with the determined ID_{90} for each bacterial pathogen. Eighteen hours later, animals will be treated either with an antibiotic, or with a bacteriophage cocktail following two distinct protocols (single i.v. bolus and infusion). Rats will be euthanized 24h after the end of placebo, phage or antibiotic administration. The cardiac vegetations and the spleen will then be removed, homogenized in saline solution and plated to determine the number of viable bacteria in tissues. We anticipate eight animals per group, i.e. a total of 96 rats for the therapeutic experiments.

Future & other potential developments

Future projects include the development of an intuitive, online platform to allow biologists and clinicians to use our methodology, including tools for the visualization and exploration of the results. Moreover, achieving a comprehensive view of phage-bacteria infection networks will also eventually provide fertile ground to network modeling to better select phage cocktails given a bacterial infection. If needed, the algorithms behind our methodology would be optimized in terms of the required computational effort. Linked to this platform, we also plan on building an open-access database of annotated interactions containing both experimentally validated (spot tests, rodent models,...) and *in silico* predicted phage-bacteria interactions, containing our own data and community-supplied data.

Being able to predict phage-bacteria interactions will also set the stage to better characterize lysins, the phage-encoded enzymes that actually digest the bacteria. Indeed, after identifying phages of interest against a bacterium using our methodology, we will be able to proceed with isolation and purification of their lysins and use them to specifically kill these bacteria using the lysins as external drugs. Our methodology will thus certainly contribute to speeding up the identification and production of lysins, which, considering their modular structure usually consisting of evolutionarily-conserved catalytic domains and a bacteria-specific cell-wall binding domain (Fischetti 2008), makes them very attractive to synthetic production as potent, narrow-host-range, antimicrobial agents against Gram-positive bacteria. Moreover, a very recent paper described for the first time clinically relevant lysins active against a Gram-negative pathogen, namely *A. baumannii* (Lood et al. 2015). Therefore, we already might get very interesting data from the present study regarding anti-*A. baumannii* lysins.

In order to further unravel the complexity behind phage biology, we also envision to develop (and improve existing) methodologies to predict other phage and phage-bacteria features (e.g. phage endotoxins, phage lifestyle depending on the bacterial target, phage-encoded mechanisms to escape bacterial defenses,...), thereby enabling faster, systematic and automated phage characterization. Together with our *in silico* prediction models for phage-bacteria interactions, this platform could ultimately be used to rapidly construct phage genomes with desired therapeutic characteristics *in vitro*, which is already being explored at the Craig Venter Institute (La Jolla, California, US) thanks to the advent of RNA-guided *in vitro* engineering.

2.4 Schedule and Milestones

Schedule, milestones and dissemination

The approximate schedule of the project is shown below. Milestones of the project include (M1) completion of a comprehensive annotated and sequenced phage library against *A. baumannii, E. coli, P. aeruginosa* and *K. pneumoniae*; (M2) development of a highly predictive *in silico* methodology to predict phage-bacteria interactions; (M3) development of *in silico* explanatory models to gain insight into the biology underlying phage-bacteria interactions; (M4) *in vitro* assays validation of the selected phages for the prospective studies; (M5) validation of the methodology using a model of infected *Galleria mellonella*; (M6) validation of the methodology using a mouse model of bacteremia; (M7) validation of the methodology using more sophisticated *in vitro* and *in vivo* models of experimental endocarditis.

Our findings will be published in peer-reviewed high-impact international journals, deposited on institutional archives for open access, and presented at international conferences. All the resources developed within this project will also be made available online on Swiss-based institutional websites.

Years	**Y1**				**Y2**				**Y3**			
Quarters	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
A.1. Sequencing of 200 bacterial strains	🟩	🟩										
A.2. Collection and sequencing of 160 phages	🟩	🟩	🟩	🟩 M1								
B.1. Dataset construction with existing data	🟦											
B.1. Feature selection		🟦	🟦									
B.1. Model selection			🟦	🟦								
B.1. Model assessment and performance				🟦	🟦 M2	🟦	🟦	🟦	🟦	🟦	🟦	🟦
B.2. Selection of explanatory models				🟦	🟦	🟦	🟦 M3					
C.1. Construct phage-bacterium infection network						🟧						
C.1. Selection of monophages & phage cocktails						🟧						
C.1. Validate selected phages with in vitro assays							🟧 M4					

C.2 *Galleria mellonella* model of infection							▨	▨ M5				
C.3.1 Mouse model of bacteremia									▨	▨	▨ M6	
C.3.2.a *in vitro* model of experimental endocarditis										▨	▨	▨
C.3.2.b *in vivo* model of experimental endocarditis										▨	▨	▨ M7
Management and dissemination	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

Risks of the project and backup plans

A. The collected phages may be redundant and we may not reach 160 different phages within one year. → We will keep collecting phages for longer to ensure that a comprehensive phage library is constructed.

B. The predictive power of the models may be lower than expected. → We will investigate further feature selection and increase the datasets with newly collected data to further improve the models.

C. The prospective studies may be negative or non-conclusive. → We will need to assess to what extent the models can be used in a translational context. In particular, this may call for modeling the dynamics of phage-bacteria infection networks to optimize the selection of phage cocktails.

Tasks of the employed staff

Employed staff will consist of (i) a junior engineer in computer science who will mainly work on modelling methodologies, model-based assessment, implementation and testing to ensure high predictive power (3 years at HEIG-VD); (ii) a senior postdoc in computational biology who will be in charge of dataset construction, feature selection and explanatory models (60% over the 3 years at HEIG-VD); and (iii) a lab technician to perform the experiments (3 years at UNIL/CHUV).

2.5 Importance and Impact

The lack of methodologies to predict phage-bacteria interactions and quickly assess their potential clinical relevance against a bacterial infection has sadly been missing in the community, slowing down research in the field. Inspired by the high-throughput *in silico* screening strategies deployed by the pharma industry for drug development, we expect our methodology to foster the time- and cost-effective use of phages in numerous fields, by enabling rapid *in silico* screening of potential candidates.

Being able to predict phage-bacteria infection networks will be of great relevance far beyond the fundamental level, where unraveling genomic features determining phage-bacteria interactions will already greatly contribute to better understanding phage biology and ecology. Our methodology could indeed be used for a broad range of applications, reflecting the urgent need for such tools in the scientific community:

- *Public health.* According to the WHO[6], antimicrobial resistance "is an increasingly serious threat to global public health that requires action across all government sectors and society". In the race for solutions against multi drug resistant bacteria, phages appear as promising candidates. Our methodology will pave the way towards personalized phage therapy.
- *Water safety*. Vinay and co-workers have recently prototyped a biosensor to specifically detect enteric bacteria from water, by fusing a green-fluorescent reporter protein to a bacteriophage targeting the bacteria (Vinay et al. 2015). In this context, our methodology could be used to cost- and time-effectively identify new phage candidates highly specific to any bacteria of interest.
- *Food industry*. Some fermentation processes in the food industry are highly dependent on bacteria. Phage infections can therefore become problematic, since they lower bacterial numbers and bacterial efficiency (Labrie et al. 2010). Our prediction tool could therefore be used to test phage resistance of *in silico* modified bacteria, thereby allowing for efficient engineering of more resistant bacterial strains.

References

Abbasifar et al. 2014.*Archives of Virology*, *159*(9), 2253–61.

Bandyopadhaya et al. 2012. *PLoS Pathogens*, *8*(11), e1003024.

Beckett & Williams 2013. *Interface Focus*, *3*(6), 20130033.

Ben-Hur & Noble 2006. *BMC Bioinformatics*, *7 Suppl 1*(Suppl 1), S2.

Coelho et al. 2014. *BMC Systems Biology*, *8*(1), 24.

Dell et al. 2002. *ILAR Journal, 43*(4), 207–13.

Denamur & Matic 2006. *Molecular Microbiology*, *60*(4), 820–7.

Dyer et al. 2007. *Bioinformatics*, *23*(13), i159–66.

Edgar et al. 2012. *Appl Environ Microbiol*, *78*(3), 744–751.

Eggimann et al. 2015. *Forum Médical Suisse*, *15*(6), 124–128.

Entenza et al. 2009. *Antimicrobial Agents and Chemotherapy*, *53*(9), 3635–41.

Fischetti 2008. *Current Opinion in Microbiology*, *11*(5), 393–400.

Flores et al. 2011. *PNAS*, *108*(28), E288–97.

Fluckiger et al. 1994. *Antimicrobial Agents and Chemotherapy*, *38*(12), 2846–9.

Fournier et al. 2015. *Burns*, *41*(5), 956–68.

Garneau et al. 2010. *Nature*, *468*(7320), 67–71.

Harris et al. 2004. *Nucleic Acids Research*, *32*(Database issue), D258–61.

Hastie et al. 2009. *The elements of statistical learning. Springer Series in Statistics* (Second ed).

Héraïef & Freedman 1982. *Infection and Immunity*, *37*(1), 127–31.

Khalifa et al. 2015. *Appl Environ Microbiol, 81(8)*:2696-705.

Kesarwani et al. 2011. *PLoS Pathogens*, *7*(8), e1002192.

[6] http://www.who.int/mediacentre/factsheets/fs194/en/

Labrie et al. 2010. *Nature Reviews. Microbiology*, *8*(5), 317–327.

Larrañaga et al. 2006. *Briefings in Bioinformatics*, *7*(1), 86–112.

Loc-Carrillo & Abedon 2011. *Bacteriophage*, *1*(2), 111–114.

Lood et al. 2015. *Antimicrob Agents Chemother, 59(4)*, 1983-91

Lu & Koeris 2011. *Current Opinion in Microbiology*, *14*(5), 524–531.

Matsuzaki et al. 2005. *Journal of Infection and Chemotherapy*, *11*(5), 211–219.

McNair et al. 2012. *Bioinformatics*, *28*(5), 614–618.

Moreillon & Que 2004. *Lancet*, *363*(9403), 139–49.

Nathan & Cars 2014. *N Engl J Med*, 371:1761–1763.

Nourani et al. 2015. *Frontiers in Microbiology*, *6*, 94.

Oliveira et al. 2013. *Journal of Virology*, *87*(8), 4558–70.

Olszak et al. 2015. *Applied Microbiology and Biotechnology*, *99*(14), 6021–33.

Pak & Kasarskis 2015. *Clin Infect Dis.* pii: civ670.

Pendleton et al. 2013. *Expert Review of Anti-Infective Therapy*, *11*(3), 297–308.

Peters et al. 2015. *BMC Genomics*, *16*(1), 664.

Polikar 2009. *Scholarpedia*, *4*(1), 2776.

Que et al. 2005. *The Journal of Experimental Medicine*, *201*(10), 1627–35.

Que et al. 2011. *J Pathog, 2011*:549302.

Que et al. 2012. *Critical Care*, *16*(4), R114.

Que et al 2013. *PloS One*, *8*(12), e80140.

Que et al. 2015. *Chest*, *148*(3), 674–82.

Que et al. 2015b. *Infection, 43(2)*, 193-9.

Rakhuba et al. 2010. *Polish Journal of Microbiology*, *59*(3), 145–55.

Ram et al 2012. *PNAS*, *109*(40), 16300–5.

Reece et al. 2010. *Campbell Biology*.

Samson et al. 2013. *Nature Reviews. Microbiology*, *11*(10), 675–87.

Sanjuán et al. 2010. *Journal of Virology*, *84*(19), 9733–48.

Seed & Dennis 2009. *Antimicrobial Agents and Chemotherapy*, *53*(5), 2205–8.

Seed 2015. *PLOS Pathogens*, *11*(6), e1004847.

Snel et al. 2000. *Nucleic Acids Research*, *28*(18), 3442–4.

Vinay et al. 2015. *PloS One*, *10*(7), e0131466.

Vouillamoz et al. 2013. *International Journal of Antimicrobial Agents*, *42*(5), 416–21.

Weitz et al. 2013. *Trends in Microbiology*, *21*(2), 82–91.

Yan et al. 2015. *Annals of Surgery*, *261*(4), 781–792.

Yellaboina et al. 2011. *Nucleic Acids Research*, *39*(Database issue), D730–5.

Yosef et al. 2014. *Bacteriophage*, *4*(1), e28491.

3. INTERDISCIPLINARY RESEARCH

This project, aiming at *in silico* prediction of phage-bacteria interactions, addresses an important limitation currently faced by biologists, clinicians, and scientists in general working with bacteriophages. It will require (i) biological expertise in microbiology to collect, sequence, annotate, and label phages and bacteria; (ii) computational expertise in machine learning and bioinformatics to select features, build the databank, create and assess the performance of predictive models; and (iii) both clinical and biological expertise in disease models and translational research to test the potential of the methodology in rodent models of infectious diseases to pave the way towards more personalized and effective phage therapy. This project is therefore interdisciplinary in essence, considering the diverse background of partners involved in this project (biologists, clinicians, computer and data scientists) as well as the various complementary expertises and state-of-the-art techniques required to successfully achieve the set goals.

Successful completion of this project will enable scientists to high-throughput screen for phage-bacteria interactions in a time- and cost-effective manner, potentially uncovering novel interactions. The pharma industry and the scientific community already benefit in numerous fields, such as drug development, of in silico tools allowing to screen for large compound libraries while ensuring time- and cost-effective experimental validations with unprecedented discovery power. Similarly, the goal of this project is to reduce the cost of screening for phage host ranges, increase discovery power, and enable the fast identification of potential phage candidates in a clinical context, where time is a very precious resource. An interdisciplinary approach is therefore both timely and necessary to tackle this challenging project. Importantly, it will also contribute to the 3R principles of replacement, reduction, and refinement in animal research while concurrently boosting new interaction-discovery power.

The computer scientists involved in this project have significant experience working with biologists and clinicians on successful interdisciplinary research projects. The partners are all located in canton Vaud, facilitating regular face-to-face meetings in addition to monthly Skype conferences to ensure good progress of the project. This will also facilitate the staff of the project to work on other partner’s premises, when required by the project. The experiments in sections A and C.2-3 of this proposal will be conducted in Lausanne in the laboratories of Dr. Yok-Ai Que (CHUV) and Dr. Grégory Resch (UNIL), who have a long-standing record of successful collaborations together and will provide the clinical and biological expertise to this project, respectively. The computational sections B and C.1 will be carried out in Yverdon in the group of Prof. Carlos Peña (HEIG-VD). All partners are committed and very enthusiastic towards working in a tight collaboration to ensure the completion and success of the project.