Searching for Bacterial Pathogens in the Digital Ocean

Paris (France)
27 - 30 September 2017
Searching for Bacterial Pathogens in the Digital Ocean

Paris, France, 27 - 30 September 2017

To be cited as:

This collection offers a broad range of titles in the marine sciences, with a particular focus on emerging issues. The Monographs do not aim to present state-of-the-art reviews; they reflect the latest thinking of researchers gathered at CIESM invitation to assess existing knowledge, confront their hypotheses and perspectives, and to identify the most interesting paths for future action.

A collection founded and edited by Frédéric Briand
CONTENTS

I – EXECUTIVE SUMMARY……………………………………………………………………………5

1. Introduction
2. Identification of useful variables / markers
3. Pathogenic traits
4. The pathogenic environment
5. Relevant, reliable, available databases
6. Statistical/ mathematical tools for data mining

II – WORKSHOP COMMUNICATIONS

Genome Complexity

- Diversity of metagenomics studies - dissecting microbiomes……………………………………27
  Mária Džunková, Francesc Peris-Bondia, Andrés Moya and Giuseppe D’Auria

- The Enhanced Infectious Disease Database (EID2) system of species interactions and location
  automatic detection and its applications…………………………………………………………33
  Maya Wardeh, Marie McIntyre and Matthew Baylis

Data Acquisition and Analysis

- Sequencing the ocean………………………………………………………………………………...45
  Chris Bowler, Frank-Oliver Glöckner and Antonio Fernandez-Guerra

- Finding Statistically Significant Patterns from Data………………………………………………53
  Mahito Sugiyama

- The “explore and exploit” strategy of bacteria: from the gut to the ocean………………………59
  Dorota Czerucka and Fernando Peruani

Genomics of Pathogenic Bacteria

- Tracking pathogens in the wild requires a plan…………………………………………………..65
  Maxime Bruto, Adèle James, Damien Piel, Sabine Chenivesse, Yannick Labreuche and Frédérique Le
  Roux
- Environmental stress and the control of gene expression in pathogenic bacteria......................69
  Charles J. Dorman

- The management and utility of whole genome sequence data for infectious disease surveillance and intervention: from hospitals, to fish-farms, to the oceans..........................................................77
  Nicola Coyle, Sion C. Bayliss, Harry A. Thorpe and Edward J Feil

- New potential NGS-based markers for detection of Vibrionaceae in marine environments.........83
  Aleksei Korzhenkov, Bogdan Efimenko, Stepan Toshchakov

Ecological Traits of Pathogenic Bacteria

- Investigating the macroecology of emerging Vibrio pathogens in the ocean using the Continuous Plankton Recorder technology.................................................................91
  Luigi Vezzulli, William H. Wilson, Carla Pruzzo

- Diversity and distribution of potential human pathogenic bacteria in the seas: novel insights from exploration of NGS databases.................................................................99
  Marc Troussellier, Jean-Christophe Auguet and Arthur Escalas

- Next generation sequencing-based approaches to characterize microbial pathogenic community and their potential relation to the Black Sea ecosystem status...........................................115
  Elena Stoica, Mariia Pavlovska, Evgen Dykyi, Konstantinos Kormas

III – BIBLIOGRAPHIC REFERENCES..................................................................................125

IV – LIST OF PARTICIPANTS.........................................................................................155
I – EXECUTIVE SUMMARY

This synthesis, sketched during the course of the workshop proper, was developed in the months thereafter on the basis of written contributions provided by most participants under Laura Giuliano’s coordination. Frédéric Briand reviewed and edited the entire Monograph. Céline Barrier was responsible for the physical production of the volume.

1. INTRODUCTION

Reports of infectious diseases affecting humans and marine organisms are more and more frequent. Whether these increases reflect better reporting or actual global trends is a subject of active research, which is receiving much attention worldwide, given the heightened human dependence on marine environments. Bacterial infections, in particular, figure among the emerging threats to human health, especially in heavily polluted coastal areas where they are associated with recreational and commercial uses of marine resources (Tamplin, 2001).

In addition evidence is mounting of an increased sensitivity of various marine organisms to infectious agents, leading to the occurrence of opportunistic pathogens, and/or amplification of resident infectious agents. Various research studies, for example, refer to ecologically and economically important species from the oceans, such as oysters and corals, which have been affected by large-scale epidemics. Increasing human pressure on marine ecosystems and ongoing climate change / warming are widely believed to further foster the spread of pathogens in the sea (Vezzulli et al., 2016).

Environmental and climatic conditions usually play a major role in the distribution of indigenous pathogenic microorganisms in the marine environment and in their transmission to humans and animals, with less certainty in the case of Vibrio cholerae (Domman et al., 2017). Beyond the (re-)emergence of indigenous pathogens (e.g. Vibrio spp.), the introduction of allochtonous pathogens by agricultural and urban runoff, ballast water or animal transfer, can be the cause of new infectious diseases with severity depending on the virulence, ecology and survival of the infectious agent.

Large-scale intensive aquaculture practices further contribute to the dramatic increase in severe disease outbreaks caused by a diverse range of pathogens, including parasites, viruses and bacteria (Soto-Rodriguez et al., 2015; Sundberg et al., 2016).

Today, ‘omics approaches (metagenomics, metatranscriptomics, metaproteomics) provide great promise for a better understanding of marine microbial communities, including marine pathogens (Coyle et al., this volume). By analysing the whole environmental DNA (eDNA), potential pathogens can be monitored for early detection and management. Novel next generation sequencing techniques allow the detection of DNA sequences even at very low concentrations, and through existing reference databases the sequences can be used to identify the presence of pathogenic microorganisms, their

1 to be cited as :

pathogenic potential, and mechanisms of evolution. Third generation, portable real-time sequencing devices are now available for genome sequencing of bacterial strains in the field (Bleidorn, 2016).

With billions of ‘omics data already available in the public repositories (see sections below) most pathogenic microorganisms will be discovered and characterized in the future by the analysis of sequences, with an ongoing shift from molecular barcoding towards metagenomics and metatranscriptomics (Pallen, 2016). No less than 40 million novel genes were predicted from the recent Tara Oceans expedition alone (Suganawa et al., 2015), and yet the molecular mechanisms of virulence of many environmental pathogens remain unknown. Even if recent metagenomic analyses have revealed that putative virulence genes are widespread in the ocean, drawing conclusions about the role of virulence genes in the absence of a model of pathogenesis would be highly premature. Caution in the interpretation of such data is strongly recommended.

In any case, all trends are leading to a “digital ocean era”. The large amount of digital information (e.g. genetic sequences in digital format) requires the development of new analytical tools to transform the huge amount of data into biological knowledge. Environmental bioinformatics will provide new solutions for this vibrant and exciting field of research, allowing to scale-up from the analysis of the thousands of marine genomes to the millions of metagenomes in their environmental context (see more in Coyle et al., this volume).

In opening the meeting, Drs Frédéric Briand and Laura Giuliano, respectively Director General and Scientific Director of CIESM, presented the overall background and objectives of the workshop to the participants (see list at the end of volume), emphasizing the urgent need for a more complete understanding of the emergence of bacterial pathogens outbreaks and the possible role of marine ecosystems as reservoirs. In summarizing the huge complexity of the marine environment, (e.g. with respect to scales, particles density, structure, etc.), they stressed the importance of tracking the phylo-geography of bacterial pathogens of humans and animals in marine areas.

The central question that framed the discussions was ‘to what extent can the digital ocean teach us something about bacterial pathogens?’ As reflected in Figure 1, which attempts to capture the many processes surrounding the analysis and mining of large data sets (oceanographic, environmental, biomedical, epidemiological, etc.), this is a complex, intricate question. In order to reveal meaningful patterns of potential marine pathogens (i.e. propagation, interaction with the host, effector delivery etc.), attention must be given to large gene expression data sets as depositories of module molecular markers of pathogenicity, and to the development of recent machine-learning algorithms that will help recognise pathogen-specific fingerprints. In fact, again and again the present volume will highlight the huge power of fast growing data sets as molecular epidemiological tools for reconstructing individual transmission events and for tracking the emergence and spread of resistant and/ or virulent clones.

To explore these questions, some fifteen experts of various geographic horizons and backgrounds (marine microbial ecology, pathogenicity and its genomic signature, virulence genes, integrated genomics and post-genomics approaches, computational molecular biology, metadata management) were invited by CIESM at the Oceanographic Institute in Paris, in late September 2017.

The present summary synthesizes the outcome of the discussions and exchanges conducted both during and in the immediate aftermath of this CIESM brainstorming workshop. Considering the fast-increasing number of datasets and platforms, and the diverse available tools for integrated analyses of metadata informing on pathogens and the marine environment, this chapter reflects the complexity of such metadata depositories, and the current difficulty to clearly identify pathogens by using the known associated (molecular) traits, in particular with respect to virulence which usually involves the interaction of different genes and gene regulators, in addition to possible host effects. The conclusions, while preliminary, are quite new in this field.
2. IDENTIFICATION OF USEFUL VARIABLES / MARKERS

In the mind of the public, bacteria can be divided into the 'good' and the 'bad', with the bad bacteria equating to pathogens. The potential for a microorganism to cause an infectious disease is influenced by several factors. A bacterium possessing genes that encode known pathogenic traits will have an increased probability to be pathogenic, but realising its potential for harm will also depend on environmental factors and on the physiological state of the host. The condition of the host, i.e. its ability to resist infection, may broaden the range of bacteria to which that host will be susceptible. In extreme cases, such as that of an immuno-compromised host, even normally benign ('good') bacteria may present a significant threat to health. For all these reasons, the development of rapid diagnostic methods with suitable sensitivity and specificity is very complex and requires multi-disciplinary, integrated efforts (Bruto et al., this volume).

Pathogenic bacteria usually maintain their virulence genes under some form of control at the level of expression to avoid wasteful production of traits that impose a fitness cost on the organism if an advantage is not achieved. In general, two strategies can be observed at work in a bacterial population. The first involves a stereotypical response in which all, or almost all, the bacteria respond to a signal that a suitable host is present by activating their virulence genes. The second involves the operation of randomly-acting genetic switches that switch virulence genes on or off in a stochastic manner across the bacterial population. In this way, just a subset of the bacteria will be primed to infect if a host appears. In practice, both strategies can be detected in individual pathogens. These strategies imply that bacteria possess the machinery to interpret their environments and to respond accordingly. Are environmental sensing systems also virulence traits? It could be argued that without them virulence mechanisms would lack ‘intelligence’ and be much less effective. Which illustrates the importance of looking beyond classical virulence genes when making assessments of the pathogenic potential of bacteria (Dorman, 1994).
The genetic switches responsible for controlling the expression of bacterial virulence genes can be indistinguishable from those that control housekeeping genes, but there are exceptions that are frequently associated with genes involved in disease. These are the bi-phasic switches that govern gene expression stochastically through mechanisms that involve DNA segment inversions, RNA methylation and point mutations within poly-pyrimidine tracts (Dorman and Bogue, 2016). It is thought that this phase-variable method of gene expression allows pathogens to evade the host defences by presenting the immune system with novel antigens that it is not primed to recognise (Moxon et al., 2006). Each of these systems has characteristic features that can be identified by interrogating bacterial genomes. For example, DNA inversion requires a site-specific recombination system that operates on inverted repeat sequences; DNA methylation by the Dam methylase requires GATC sequences that are associated with transcription signals and poly-pyrimidine tracts produce frame-shift mutations in open reading frames that alter the repertoire of surface proteins expressed by pathogens. All of these features are discoverable by genome sequence analysis (Chen et al., 2014). Typically, the genes affected encode secreted proteins or cell surface components and will have protein secretion signals, which are also detectable bioinformatically.

Gene regulation using 'conventional' transcription factors is common among virulence genes that are subject to stereotypic control. These DNA binding proteins are divided into classes based on their domain structure and can be studied bioinformatically. Examples include the AraC-like proteins, the LysR-like proteins and members of the sensor-kinase/response regulator superfamily (Maddocks and Oynston, 2008; Yang et al., 2011). Almost every major virulence gene control network contains examples of these transcription-controlling proteins. In many cases the binding sites of the proteins can also be detected bioinformatically. The proteins have signal reception domains that bind small molecules or metals; the response regulators are phosphorylated by the sensor kinases on conserved aspartic acid residues, making the proteins proficient for DNA binding (Dorman and Dorman, 2017).

DNA base composition can be a clue to the presence of virulence genes in the genome. Abnormal A+T base content may indicate that the genes in question have arrived by horizontal transfer and have been maintained because they added some advantageous features (possibly connected with pathogenicity) to the receptor organism. The link is quite reliable based on the studies done so far. While not every horizontally-acquired gene is a virulence gene, most virulence genes have been acquired horizontally (Gyles and Boerlin, 2014). These unusual DNA structural properties have implications for gene regulation, the choice of sigma factor to operate RNA polymerase and the likelihood that the genes will be targeted by transcription silencing nucleoid-associated proteins. Other clues that a portion of the genome has been acquired by lateral transfer include the presence nearby of phage attachment sites and/or genes (or pseudogenes) encoding integrases or transposases.

The expression of many virulence genes is controlled by small RNAs (sRNA) and the sRNA-mRNA interaction frequently requires a chaperone protein (Holmqvist et al., 2016). Small RNAs can be discovered by whole genome analysis methods, as can potential RNA chaperone proteins (Barquist et al., 2016). RNA control that acts in cis can also be predicted bioinformatically because it frequently relies on differential transcription termination that leads to alternative folding of the transcript.

The creation of physiological variety across a bacterial population is an excellent strategy for survival because it reduces the risk that all members of the population will be eliminated by a single catastrophe. The arrival of a powerful antibiotic in the midst of a population of susceptible bacteria is an example of such a catastrophe. However, if the action of the antibiotic is restricted to periods when the bacterium is growing, any non-growing organisms will escape death or inhibition. Persistence describes the ability of bacterial populations to bring forward metabolically inert members in a completely stochastic manner. These non-growing organisms are temporarily non-susceptible to antibiotic action and will be capable of carrying on the life of the population after the antibiotic is removed. Persistence mechanisms often involve a 'toxin/anti-toxin' binary system in which the toxin component inhibits growth following the stochastic disappearance of the (usually-unstable) anti-toxin (Harms et al., 2016). The presence of such systems can be an aid to virulence and may underlie recurrent infections that appear to be recalcitrant to antibiotic therapy. Persistence is distinct from antibiotic resistance, which involves mechanisms by which the bacterium becomes permanently
resistant through modification of the drug target, over-expression of the target, inactivation of the drug, or an ability to pump the drug out of the cell. Nevertheless, the detection of resistance genes in association with classical virulence traits may be indicative of enhanced pathogenic potential. Similarly, the presence of persistence systems could indicate an ability on the part of the bacterium to outwit host defences that rely on killing metabolically active bacteria (Maisonneuve and Gerdes, 2014).

Each infection is characterised by a minimum infectious dose. The fewer organisms needed to initiate a successful infection in a healthy host, the more virulent that pathogen will be. Do bacteria ‘count’ one another to determine if they have reached a threshold at which initiating an infection is worthwhile? Data from studies of bacterial cell-to-cell signalling suggest that they do (Papenfort and Bassler, 2016). Signal production and signal detection form the basis of quorum sensing. Here, small molecules are produced by individual bacteria in concentrations too low to elicit a response, but when pooled with the same molecules from other nearby bacteria a threshold is crossed that leads to the elaboration of a new behaviour, perhaps directly involved in the infection of a host. The genes that encode the enzymes of the pathways for expression of the signalling molecules and the genes that encode the receptors that detect those molecules can all contribute to the pathogenic potential of the bacterium.

3. PATHOGENIC TRAITS

Bacteria often engage on pathogenic behaviour in order to win new resources or to escape from an unfavourable environment. As to marine environments, depending on the environmental conditions, bacteria can move in a free individual manner or remain in the same place to form colony groups and colonize surfaces. As a group, bacteria can optimize growth and survival by the presence of different cell types that are able to perform specialized functions (i.e. better access to nutrients; better defence mechanisms for protection against unfavourable environmental conditions, etc.). Some bacteria can secrete polysaccharides to form biofilms which enhance adhesion, survival, and movement. The main pathogenic factors associated with such functions are surface polysaccharides (capsule, lipopolysaccharide, and glucan), S-layers, iron-binding systems, exotoxins and extracellular enzymes, secretion systems, fimbriae and other nonfilamentous adhesins, motility and flagella. An integrated compilation of bacterial pathogenic traits in *Aeromonas* spp. is provided by Tomas (2012).

Many virulence genes are switched on in response to a shortage of iron. Although this is the most abundant metal on earth, iron is largely unavailable to biology and is strongly sequestered by living organisms. Bacteria must therefore compete with their hosts to acquire it. Many successful bacteria manufacture iron-carrying molecules (siderophores) that have a higher affinity for iron than those of their customary hosts. They also have efficient transport systems to bring the iron-siderophore complexes into their cytoplasm where the metal can be released and used, for example to create active centres in proteins involved in electron transport in the respiratory chain (Palmer and Skaar, 2016). Iron acquisition genes can be found in bacterial chromosomes but they are also found on plasmids, including plasmids that harbour virulence genes. Detection of iron uptake genes does not in itself prove that a bacterium is a pathogen but their association with other disease-associated genes may reveal pathogenic potential.

According to the scientific literature, temperature increases in oceans can stimulate opportunistic pathogens and favour waterborne disease outbreaks, including sometimes the “reverse zoonosis”, namely the transmission of human pathogens to marine organisms. Among many examples figure various coral disease outbreaks (Ben-Haim *et al.*, 2003; Sutherland *et al.*, 2011), which have been also observed in Mediterranean waters (Rubio-Portillo *et al.*, 2014).

3.1 Genes / proteins with clinical relevance

Secretion systems are required for the export of virulence factors, including toxins (Mecsas and Strauss, 1996). Toxin genes can be found in isolation or as part of operons that include genes for toxin subunits or other virulence traits. Adhesins are essential for the colonisation of surfaces and for
assembling the bacteria into communities. Matrix-binding proteins allow bacteria to adhere tightly to the surfaces of host cells on the exterior or in the interior of the body (Chagnot et al., 2012). Biofilm protects the bacterial community from the environment – aids colonisation of rocks in watercourses, on the seashore, the linings of pipes in the plumbing in buildings, the surface of teeth, etc. Biofilm glues microbes together and helps them resist shear forces or other forces that might wash them away. In some cases, flagellar motion provides bacteria with the capacity to avoid being trapped by the host cell physical defences (i.e., epithelial mucus in the gastrointestinal tract; see Czerucka and Peruani, this volume). Capsules mask bacterial surface antigens and help the bacteria to evade the host immune system. Capsules also defend the bacteria from desiccation and from thermal stress. Colanic acid is produced by many bacteria in response to low temperature and osmotic stress. The drying action associated with osmotic stress results in bacterial accumulation of compatible solutes that replaces lost water. All of this relies on a system for sensing and responding to osmotic stress. The list of known virulence factors of *V. parahaemolyticus* is provided below as an example (Table 1).

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Activity</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDH</td>
<td>Thermostable directed hemolysin</td>
<td>Pore forming toxin</td>
<td>Cytotoxicity &amp; enterotoxicity</td>
</tr>
<tr>
<td>TRH</td>
<td>TDH related hemolysin</td>
<td>Pore forming toxin</td>
<td>Cytotoxicity &amp; enterotoxicity</td>
</tr>
<tr>
<td>MAM7</td>
<td>mce domain</td>
<td>Binds to fibronectin &amp; phospholipid / atidic acid</td>
<td>Attachment to the host cell</td>
</tr>
<tr>
<td>VopQVP1680</td>
<td>Non conserved</td>
<td>Binds to V-ATPase</td>
<td>Autophagy induction</td>
</tr>
<tr>
<td>VopSVP1686</td>
<td>Fic domain</td>
<td>AMPylates Rho family GTPases</td>
<td>Cytoskeleton disruption</td>
</tr>
<tr>
<td>VPA450</td>
<td>Inositol 5-phosphate</td>
<td>Hydrolyzes PI(4,5)P2 to PIP4</td>
<td>Plasma &amp; membrane disruption</td>
</tr>
<tr>
<td>VopCVPA1321</td>
<td>Rac &amp; CDC42 deamination</td>
<td>Cytotoxic necrozing factor</td>
<td>Disturbs actin, bact. invasion</td>
</tr>
<tr>
<td>VopTVPA1346</td>
<td>ADP/ribosyltransfase</td>
<td>ADP-ribosylates Ras</td>
<td>Unknown</td>
</tr>
<tr>
<td>VopA/T</td>
<td>Acetyltransferase</td>
<td>Inhibits MAPK signaling</td>
<td>Imm-resp.suppress.</td>
</tr>
<tr>
<td>VopVVPA1357</td>
<td>Non-conserved</td>
<td>Actin binding / bundling</td>
<td>Cytotoxicity &amp; enterotoxicity</td>
</tr>
<tr>
<td>VopLYPA1370</td>
<td>WH2 domain</td>
<td>Actin nucleation</td>
<td>Induction of actin stress fiber</td>
</tr>
</tbody>
</table>
3.2 Quorum sensing

What are the genes involved in the synthesis of the quorum-sensing molecule, the degradation of the quorum sensing molecule and the detection of such molecules? Are the molecules confined to intra-species signalling or can they participate in inter-species communication? What behaviour might be controlled by the signal? Some influence cell-to-cell transfer of DNA, others control the expression of biofilm material, while others regulate the expression of bioluminescence. Many bacteria have integrated their virulence genes into control networks that have a role in quorum sensing (e.g. *Pseudomonas*).

3.3 Whole metabolic pathways

The fewer complete metabolic pathways a microbe has, the more dependent it is on its host or other environments for survival. *Mycoplasmas* spp. have a very small genome and are obligate parasites of humans and other animals. *Mycobacterium tuberculosis* has many genome gaps that have occurred over the period it has lived in intimate association with cattle and humans. Virulence factors are often produced in branches of central metabolic pathways. For example, many siderophores contain a ring structure that is derived from chorismate, an intermediate in the same pathway that produces para-amino benzoic acid (pABA), folate and the aromatic amino acids.

3.4 Invasion factors

Invasion factors come in many forms but share the property of being able to get a normally non-phagocytic cell to phagocytose the microbe. Many systems inject effector proteins into the cytosol of the host cell to modify its cytoskeleton so that the surface envelops the microbe, bringing it inside. Pathogens can also use specialist proteins to escape from phagocytic vacuoles and to recruit host actin for locomotion across a cell and through the barriers that separate cells. They can weaken the structures that hold host epithelial cells together in tissues, facilitating the passage of the bacteria between those cells. Some pathogens (*Salmonella*) can subvert host defence cells such as macrophage, using them as a means to move between tissues in the host (Liss and Hensel, 2015). Bacteria also have the potential to induce cell death in macrophage, killing them and releasing the pathogens (Sridharan and Upton, 2014).

3.5 Competitive fitness

Competitive fitness is a very important concept in bacterial biology (von Bronk et al., 2017). It compares the relative abilities of different bacteria to reproduce in a given environment. Any change to the genetic composition of the bacterium that enhances fitness may improve its virulence too, although there are often tradeoffs. For example, acquiring a plasmid that encodes a highly efficient iron uptake system may benefit competitive fitness in a low iron environment. In an anaerobic environment where iron, in its reduced Fe$^{2+}$ state, can enter the cell without the aid of an uptake system, the burden of plasmid carriage may undermine the fitness of the bacterium. Fitness is enhanced by carrying just enough genetic capacity to survive and reproduce while ensuring that the expression of that genetic capacity is regulated very finely to ensure the optimal use of energy and resources. Sluggish responses to environmental change or wasteful operation of metabolic pathways can severely undermine fitness; so can inappropriate expression of virulence traits. Failing to avoid or evade host defences or other environmental hazards due to poor or inaccurate interpretation of the environment can prove deleterious to the bacterium. The list of virulence determinants for *Vibrio* species is provided here as an example (Table 2).
Table 2. Examples of virulence determinants for Vibrio species. 

- **cheR**: Chemotaxis gene; 
- **Fla, Laf**: Flagellar & lateral flagellar gene; 
- **GacA/s**: two-component signal transduction system; 
- **PilA**: Type IV pilin; 
- **PilD**: Pre pilin peptidase; 
- **CpsA**: capsular polysaccharide A; 
- **GbpA**: GlcNac-binding protein A; 
- **MAM**: Multivalent Adhesion molecule; 
- **Wza**: exopolysaccharide genes; 
- **Orf1**: insertion site of mini-Tn5phoA transposon; 
- **OmpU**: outer membrane protein; 
- **Apha/OpaR**: transcription regulator; 
- **T3SS1/2**: type three secretion system 1 or 2; 
- **T6SS**: Type six secretion system; 
- **tdh**: thermostable direct hemolysin; 
- **trh**: thermostable direct hemolysin (tdh)-related hemolysin; 
- **hlyIII, VvhA, HlyA**: hemolysin in different species; 
- **CTX**: cholera toxin; 
- **toxR**: toxin regulated system; 
- **tcp**: toxin co-regulated pilus; 
- **RTX**: exotoxin and virulence factors.

<table>
<thead>
<tr>
<th>Category</th>
<th>Fitness factors</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motility</td>
<td>cheR</td>
<td>V. anguillarum</td>
</tr>
<tr>
<td>Flagella</td>
<td>Fla, Laf</td>
<td>V. parahaemolyticus</td>
</tr>
<tr>
<td>Biofilm formation</td>
<td>GacA/S, pilA, pilD</td>
<td>V. vulnificus</td>
</tr>
<tr>
<td>Capsule</td>
<td>CpsA</td>
<td>V. parahaemolyticus</td>
</tr>
<tr>
<td>Adhesion</td>
<td>GbpA</td>
<td>V. cholera</td>
</tr>
<tr>
<td>Chitinase</td>
<td>V. anguillarum</td>
<td></td>
</tr>
<tr>
<td>MAM</td>
<td>V. parahaemolyticus</td>
<td></td>
</tr>
<tr>
<td>wza, orf1</td>
<td>V. anguillarum</td>
<td></td>
</tr>
<tr>
<td>OmpU</td>
<td>V. vulnificus, V. fisheri</td>
<td></td>
</tr>
<tr>
<td>Quorum sensing</td>
<td>Apha</td>
<td>V. parahaemolyticus</td>
</tr>
<tr>
<td>OpaR</td>
<td>V. parahaemolyticus</td>
<td></td>
</tr>
<tr>
<td>Secretion</td>
<td>T3SS1</td>
<td>V. parahaemolyticus, V. cholera</td>
</tr>
<tr>
<td>T3SS2</td>
<td>V. parahaemolyticus</td>
<td></td>
</tr>
<tr>
<td>T6SS</td>
<td>V. parahaemolyticus</td>
<td></td>
</tr>
<tr>
<td>Hemolysis</td>
<td>tdh</td>
<td>V. parahaemolyticus</td>
</tr>
<tr>
<td>trh</td>
<td>V. parahaemolyticus</td>
<td></td>
</tr>
<tr>
<td>hlyIII</td>
<td>V. vulnificus</td>
<td></td>
</tr>
<tr>
<td>VvhA</td>
<td>V. vulnificus</td>
<td></td>
</tr>
<tr>
<td>hlyA</td>
<td>V. cholera</td>
<td></td>
</tr>
<tr>
<td>HlyA</td>
<td>V. tubiashii</td>
<td></td>
</tr>
<tr>
<td>Toxicity</td>
<td>CTX, toxR, tcp</td>
<td>V. cholera</td>
</tr>
<tr>
<td>RTX</td>
<td>V. vulnificus</td>
<td></td>
</tr>
</tbody>
</table>

3.6 Bacteriophages, integrons and mobile genetic elements

Mobile genetic elements of all kinds drive bacterial evolution over impressively short timescales (Koonin and Makarova, 2017; Wu et al., 2015). They allow bacteria to sample a very wide range of genetic elements and to experiment with them. Incorporating useful ones permanently into the genome permits a bacterium to acquire sophisticated new behaviours in one step (for example, the ability to resist an antibiotic; the ability to utilise a complex carbon source that was previously beyond its physiological capacity). Many overt virulence systems such as cholera toxin (Waldor and Mekalanos, 1996) are delivered by bacteriophages. The process of lysogenic conversion’ in which a bacteriophage interrupts one virulence gene physically while simultaneously delivering a new one creates variety among the virulence traits of a population of pathogens (see more in Dorman, this volume). Integrons are a means of building a repertoire of antibiotic resistance cassettes in the genome (Gillings, 2017). These collections can themselves become mobile, spreading the multi-drug-resistance phenotype through the microbial population. Plasmids harbour a wide variety of traits that are relevant to
virulence. They can be self-transmissible between bacterial cells or capable of being mobilised (Bañuelos-Vazquez et al., 2017). Plasmids carry regions of DNA sequence homology with the bacterial chromosome, allowing them to become fused with the chromosome and to mediate its rearrangement and evolution, a behaviour that has important implications for the fitness of the bacterium and its ability to interact with the host (Humphrey et al., 2012). Plasmid transfer rates are intimately linked to the formation of biofilm, which is itself linked to virulence. Plasmids often carry transposons and these in turn carry genes for resistance to antibiotics and heavy metals. Transposition is a key driver of genome evolution, not only causing genes to be knocked out but also causing previously silent genes to become active. Plasmids vary greatly in size: the so-called second chromosome in *Vibrio cholerae* is really a very big plasmid (Orlova et al., 2017). This serves to illustrate the capacity of plasmids to absorb more and more genetic information. By being separate from the chromosome that houses the genes essential for the survival of the bacterium, plasmids can be jettisoned when selective pressure for their maintenance is absent. In practice, many large plasmids have evolved systems that kill/ inhibit bacteria that manage to lose the plasmid (Gerdes et al., 2005). These toxin/anti-toxin systems have been co-opted by the bacteria and serve randomly to create persister cells in the population (see above).

### 3.7 CRISPR/cas

CRISPR/cas is a form of immunity that allows a bacterium to identify and destroy the DNA of an invader, such as a bacteriophage (Horvath and Barrangou, 2010). While many phages are temperate and do not kill the bacterium (at least immediately) there are virulent phages that will hijack the bacterial cell to reproduce themselves straightaway, leading to the death of the bacterium. Being able to destroy these viruses is very useful to the bacterium. Also useful is the ability to use restriction endonucleases to destroy foreign DNA and to use DNA methylation patterns to distinguish between self and non-self at the level of DNA (Loenen et al., 2014). A failure to detect a working CRISPR/cas system in a pathogen may indicate that it is evolving at a high rate through the risky strategy of sampling a wide range of DNA from the environment.

### 3.8 Resistance

It is important to recall that perhaps 70% of antibiotics are made by bacteria and so the producers have to possess the means to survive their own products (Martin and Liras, 2012). These resistance genes can be exported via horizontal gene transfer throughout the microbial world (Yamashita et al., 2014). Loss of competitive fitness associated with the carriage of a resistance gene that is currently not under selection represents a force that limits the process. Nowadays, widespread pollution of the environment with low concentrations of antimicrobials of all kinds is imposing just this type of selective pressure and is driving the spread of resistance genes. Pathogens are dangerous; antibiotic resistant pathogens are very dangerous indeed.

### 3.9 Epigenetic markers

These are associated with certain virulence genes and operons and epigenomic studies are likely to reveal many more (Chen et al., 2014). For example, *agn43*, the gene for antigen 43, is an important autotransporter that is under dual control by Dam-mediated methylation and oxidative stress in *E. coli*. The *pap* operon that encodes Pap pili, important for pyleonephritis and kidney infection, is controlled phase-variably by Dam methylation and the leucine-responsive regulatory protein, Lrp (Hernday et al., 2003). These genetic switches are characterised by alternate fully- and hemi-methylated states that are permissive or non-permissive for gene expression, creating cell-surface variety among bacteria in a genetically homogeneous population. The result is an increased probability of evading the host defences during infection.
3.10 Pathogenic lineages

To date, the known lineages containing human and/or animal pathogens are the Flavobacteriabacteroides, the Spirochetes, the Chlamydia, the Cyanobacteria, the Proteobacteria and the Gram-positive bacteria (i.e. *Mycobacterium*, *Clostridia*, *Listeria*, *Rhodococcus* and *Streptococcus*). A large majority of known marine pathogens belong to the Gammaproteobacteria. Within these, the genus *Vibrio* alone contains 12 recognized human pathogens and many more animal pathogens. Other proteobacteria frequently associated to disease are *Aeromonas* and *Shewanella*.

However, diagnosis based on taxonomy is not sufficient to conclude about pathogenicity since the functional unit of pathogenesis is more often the strain or clone within a species thanks to recent acquisition of lateral gene transfer (LGT). There is an urgent need for accurate and rapid laboratory diagnostic methods leading to better control and treatment strategies, which shall include detection of specific virulence factors playing a role in the different kinds of infections (ex. adherence, contact-independent factors etc.)

4. THE PATHOGENIC ENVIRONMENT

Changes in host range, in pathogenic traits displayed in the same host, and the geographic distribution of a disease complex form three distinct sets of complementary and only slightly intersecting disease emergence scenarios. Together, these scenarios present the full picture and range of possible disease emergence dynamics (Engering et al., 2013). A new era in medical science has dawned with the realization of the critical role of evolutionary and ecological factors, including both the microbial community structure and host health conditions in bacterial infectious diseases.

By providing for the first time a more comprehensive view of the “microbiome”, the *Human Gut Microbiome* initiative showed the importance of selective pressures and community dynamics in shaping the microbiome (including the “extra-intestinal” one) in diseased humans and in gut pathogens. These breakthrough results have opened new scenarios whereby human ‘disease susceptibility’ could well exhibit geographical patterns depending on social, economic and ecological features (Ruth et al., 2016).

4.1 Geographic location

Geographic location determines the probability of potential contact with host organisms. Coastal regions with dense human populations obviously represent areas with higher risks of disease transmission than other more pristine or open ocean sites. Ocean circulation further determines the connectivity of different sites with each other, such that Lagrangian transport distances are more relevant in the ocean than are actual geographic distances (see CIESM 2016). Thus one will find regions where transport times are short and extensive such as the Gulfstream in the North Atlantic, and areas where mixing is extremely limited such as in the large oceanic gyres. Further the biodiversity of different oceanic regions will determine quite distinct oceanic biogeographical provinces. As for all other oceanic microbes, the spatial distribution of pathogenic organisms is driven by wind, vertical transport, transport in ballast water, or rafting on marine litter such as microplastics (see CIESM 2014).

Mechanisms responsible for genome plasticity are found to specifically drive bacterial adaptive response and bacterial evolution in each host environment (see Dorman; Vezzulli et al., both in this volume).
4.2 Physico chemical variables

These physical and chemical signals individually and collectively impose selective pressure on the bacterial metagenome, allowing some microbes to prosper while others decline. Knowledge of environmental structure and composition and of the genetic make-up of the bacterial population allows informed predictions to be made about which organisms are likely to inhabit particular niches and which are likely to be excluded. For example, a strict aerobe is likely to do poorly in a fully anoxic environment. When making predictions, one must keep in mind the ability of bacteria to become quiescent, entering a long-term dormant state, and in some cases to sporulate, before ruling out *a priori* the ability of bacteria to inhabit any part of the natural environment.

4.3 Biotic interactions

In addition to environmental (or abiotic) variables, the likelihood that microbes form myriads of associations between each other is receiving renewed interest. This has been explored in depth with the *Tara* oceans dataset by exploring global patterns of co-occurrence of organisms (Lima-Mendez *et al.*, 2015; Bowler, this volume). Moreover, a growing volume of research data supports the hypothesis that marine organisms may function as inter-epizootic reservoirs or vectors of pathogenic bacteria, and sometimes carry them over long distances (Trousselier *et al.*, this volume). A notable example is the association of *Vibrio cholerae* as a commensal microbe on marine copepods (Vezzulli *et al.*, this volume).

5. RELEVANT, RELIABLE, AVAILABLE DATABASES

The evolution of molecular biology protocols and sequencing technologies, from Sanger sequencing through 2nd and 3rd generations parallel methods, allows us to collect a continuously growing amount of sequencing data which historically had a Moore's law doubling time of seven months, now estimated for 2nd generation Illumina sequencing at about twelve months (Goodwin *et al.*, 2016). Even within this more conservative view of a twelve months doubling time, we will face such a huge amount of data within a decade that the main problem will rely more on data transfer than on proper data processing (Muir *et al.*, 2016; Stephens *et al.*, 2015).

At this point in time where sequencing is no longer an economical or a technical issue, the bottleneck for many laboratories is in the analysis step. Yet shotgun and ribosomal genes amplicons data continue growing in volume at a vertiginous rhythm. Fortunately, the current tendency is to collect and to organize sequencing data so as to provide the users with powerful, open source tools such as the Metagenomics RAST Server (see below) to analyse and compare datasets from different environments.

5.1 Sequences / Nucleotides Databases

Publication of any genomics data requires submission of the corresponding DNA sequence data to one of the three main nucleotide archives joined under the “International Nucleotide Sequence Database Collaboration (INSDC)” initiative ([http://www.insdc.org/](http://www.insdc.org/)), namely:


b/ The DNA Data Bank of Japan (DDBJ, Japan, http://www.ddbj.nig.ac.jp/);
The content of each database is automatically mirrored at least every twenty-four hours. These databases have been increasing in size since their first appearance in the late 1980s and, given the advances in sequencing technologies, they have also been growing in complexity. Nowadays, in addition to the sequencing data, they also request the deposition of as many associated metadata as possible. This has required the development of standards that must be followed in order to deposit sequence information, which is crucial for allowing heterogeneous datasets to be compared with each other. The Genomics Standards Consortium (http://gensc.org/) plays a crucial role in ensuring the application of these recommendations. For marine metagenomics studies, the recommendations described in Ten Hoopen et al. (2015) are particularly relevant.

5.2 EMG (EBI Metagenomics Portal)

(https://www.ebi.ac.uk/metagenomics/)

The EBI Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data which aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample. It is possible to freely browse all the public data in the repository, which includes *Tara* Oceans, OSD, etc. (Mitchell et al., 2017) (more details in Bowler et al.; Villarroya et al., both in this volume)

5.3 PANGAEA

(https://www.pangaea.de/)

The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. Each dataset can be identified, shared, published and cited by using a Digital Object Identifier (DOI).

5.4 OBIS

(https://www.iobis.org/)

OBIS – the Ocean Biogeographic Information System - is a global open-access data and information clearing-house on marine biodiversity for science, conservation and sustainable development. It emanates from the Census of Marine Life (2000 – 2010) and is pursued, with contributions of 500 institutions, under the aegis of the IOC International Oceanographic Data Exchange Programme (IODE).

5.5 MG-RAST

(https://metagenomics.anl.gov/)

MG-RAST is a web application server that allows the users to upload metagenomes for automated analysis and phylogenetic classification of sequence fragments and functional classification of samples.

5.6 Ribosomal sequences databases

One of the most important resources enabling the taxonomic description of microbial life is represented by databases and search tools providing the research community with aligned and annotated rRNA gene sequence data such as the Ribosomal Database Project (Cole and Tiedje, 2014 ;
Cole et al., 2005) and the SILVA databases, jointly with the ARB project for phylogenetic tree reconstructions and representations, which includes small and large ribosomal gene subunits from the Bacteria, Archaea and Eukarya domains (Quast et al., 2013; Yilmaz et al., 2014). More precisely, the RDP and SILVA projects began with less than 500 entries obtained in Charles Woese laboratory in 1991 and now contain more than three (RDP) and six (SILVA) million sequences in their most recent releases.

Interestingly, together with Bacteria/Archaea domains, Fungi have also been targeted for wide spectrum taxonomic characterization by the means of the Intergenic Transcribed Spacers (ITS) region (Deshpande et al., 2016). Today, 18S rDNA sequences from eukaryotes are most comprehensively represented in the PR2 database (Guillou et al. 2013).

5.7 MAGE (Microbial Annotation Genome and Analysis, Genoscope)

A large amount of public genomic data concerning various bacterial species have been integrated in the MicroScope platform to ease analysis by a common set of methods and parameters (e.g. the VibrioScope project benefits from dynamic and permanent updates of genome annotations). Twice a year, trainings are organised by the MAGE team for genome annotation and comparative genomics, RNAseq and metabolic network analysis.

6. STATISTICAL/MATHEMATICAL TOOLS FOR DATA MINING

Data mining is the discipline of discovering patterns of various types of variables from databases (Aggarwal, 2016; Leskovec et al., 2014; Zaki and Meira, 2014), and the techniques of machine learning plays a central role to achieve the task (Bishop, 2007; Murphy, 2012). That sector is vast and our meeting focused on recent advances of data mining and machine learning methods that can find potential associations of variables from biological data.

To find relevant variables from a dataset, feature selection is the representative approach in machine learning. Feature selection detects variables, or features, that are associated with the target variable from the set of all variables in a given dataset (Guyon and Elisseeff, 2003). The target variable can be binary (0 and 1 for cases and controls) in a case-control study or continuous in regression. The simplest method is variable ranking, where we compute the association score, for example, Pearson’s correlation coefficient, to detect linear association or the mutual information for nonlinear association, between each variable and the target variable, followed by ranking the variables according to the association scores. Then one can find top-ranked highly associated variables from the ranking. This type of the two-step procedure, measuring association and making a ranking, is called a filter method. Although this approach can be easily applied for removing unnecessary variables from a dataset, redundant features might be selected as interactions between variables are not considered. For instance, if a dataset contains exactly the same variables that have the strong association with the target variable, both variables are selected.

The other two approaches for feature selection are a wrapper method and an embedded method, where the quality of each variable is assessed by the accuracy of a prediction model with respect to the target variable. A wrapper method repeats to construct a prediction model for each subset of variables; hence it is computationally too expensive in most cases. By contrast, in an embedded method, variables are automatically selected during the process of learning a prediction model from a dataset. A popular

---

1 www.genoscope.cns.fr/agc/microscope/about/collabprojects.php?P_id=35
method is lasso (Tibshirani, 1996) that learns a linear prediction model, where a set of variables, which receive nonzero coefficients, is automatically selected in the learning process by regularizing the number of variables. The joint additive effect of selected variables maximizes the prediction accuracy of the model. The recent advances in selective inference (Taylor and Tibshirani, 2015) now enable us to assess the statistical significance of selected features in an embedded method.

To discover not single features or linear combinations of features but patterns, that is, combinations of features with multiplicative effects, pattern mining is the promising approach in feature selection (Aggarwal and Han, 2014; see more in Sugiyama, this volume), which was originally developed to find frequently co-purchased items in market basket analysis (Agrawal and Srikant, 1994) and increasingly used in a wide range of applications including bioinformatics (Zhang et al., 2014). In particular, significant pattern mining (Llinares-López et al., 2017; Terada et al., 2013) has been recently developed, where one can find all variable combinations that are statistically significantly associated with the target variable while rigorously controlling the family-wise error rate (FWER). Moreover, one can apply pattern mining to detect sequential patterns (a sequence of variables) from time-series data and substructures of graphs from a graph-structured data.

In addition to feature selection, dimension reduction is often used to reduce the number of variables in the dataset, where variables are not directly selected but transformed into principal variables. Among a number of dimension reduction techniques, t-SNE is recently becoming a popular method and often used to visualize a multi-dimensional dataset (van der Maaten and Hinton, 2008).

Clustering is the standard approach to find groups of data points, which does not usually need the target variable. Given the number $K$ of groups of data points, or clusters, the $K$-means algorithm divides data points into $K$ disjoint clusters. Since the clustering quality obtained by the $K$-means algorithm is deteriorated by outliers in a dataset, one can perform outlier detection to remove such outliers before applying clustering, which can be also used as pre-processing for feature selection. State-of-the-art algorithms can efficiently detect outliers from a dataset using subsampling of data points (Sugiyama and Borgwardt, 2013).

By combining pattern mining and clustering, text mining can be achieved to analyse large amounts of text data. For example, one can automatically categorize text data such as scientific papers or news articles by finding frequently co-occurring words from them as keywords, followed by clustering them according to such keywords. Text mining is now actively used in the construction of biological databases (Wardeh et al., 2015 and in this volume).

Recently, machine learning techniques are used not only for directly predicting the target variable but for designing experiments. Bayesian optimization offers the sequential experimental design strategy, which tells us which data point should be examined at the next experiment to efficiently maximize the prediction accuracy with respect to the target variable (Mockus, 2011). This technique has been already successfully used in various applications such as materials science (Ju et al., 2017). Thus it might be interesting to use Bayesian optimization to build a biological database in an efficient manner.

References


CIESM Workshop Monographs n°49


