Contents lists available at ScienceDirect

Virus Research

Review

A new era of virus bioinformatics

Bashar Ibrahim[a,b], Dino P. McMahon[a,c,d], Franziska Hufsky[a,b], Martin Beer[a,e], Li Deng[a,f],
Philippe Le Mercier[a,g], Massimo Palmarini[h], Volker Thiel[a,i,j], Manja Marz[a,b,*]

[a] European Virus Bioinformatics Center, Jena, Germany
[b] RNA Bioinformatics and High Throughput Analysis Jena, Friedrich Schiller University Jena, Jena, Germany
[c] Host Parasite Evolution and Ecology, Institute of Biology, Free University of Berlin, Berlin, Germany
[d] Department for Materials and Environment, BAM Federal Institute for Materials Research and Testing, Berlin, Germany
[e] Institute of Diagnostic Virology, Friedrich-Loeffler-Institute, Greifswald, Germany
[f] Institute of Virology, Helmholtz Zentrum Munich, Munich, Germany
[g] Swiss-Prot Group, SIB,CMU, University of Geneva Medical School, Geneva, Switzerland
[h] MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom
[i] Federal Department of Home Affairs, Institute of Virology and Immunology, Bern and Mittelhausen, Switzerland
[j] Department of Infectious Diseases and Pathobiology, University of Bern, Bern, Switzerland

ABSTRACT

Despite the recognized excellence of virology and bioinformatics, these two communities have interacted surprisingly sporadically, aside from some pioneering work on HIV-1 and influenza. Bringing together the expertise of bioinformaticians and virologists is crucial, since very specific but fundamental computational approaches are required for virus research, particularly in an era of big data. Collaboration between virologists and bioinformaticians is necessary to improve existing analytical tools, cloud-based systems, computational resources, data sharing approaches, new diagnostic tools, and bioinformatic training. Here, we highlight current progress and discuss potential avenues for future developments in this promising era of virus bioinformatics. We end by presenting an overview of current technologies, and by outlining some of the major challenges and advantages that bioinformatics will bring to the field of virology.

1. Crosstalk between virology and bioinformatics

Viruses are the cause of a considerable burden to human and animal health (Kirk et al., 2015). In recent years, we have witnessed both the emergence of new viral diseases (e.g. MERS, SARS; see Fig. 1) and the re-emergence of known diseases in new geographical areas (e.g. Zika, Dengue and Chikungunya). The increased global risk of viral emergence is due to a variety of social, environmental and ecological factors. Climate change, deforestation, urbanization, and the unprecedented mobility of goods, people, animals and disease vectors are all elements that are facilitating the spread of viral diseases and creating potentially ideal conditions for pandemics.

The economic burden of viral diseases is enormous. The costs of all global disasters are currently estimated at 150 billion USD per year of which 30 billion USD are attributable to infectious disease outbreaks alone.[1] Viruses can also cause diseases in animals and plants. Diseases of livestock affect food security and inflict considerable economic damage. For example, annual losses due to foot-and-mouth disease are

between 6.5 and 21 billion USD in endemic areas (Knight-Jones and Rushton, 2013).

Virologists have traditionally concentrated on studying viruses that cause disease in humans, animals or plants. However, there is a staggeringly large number of viruses in the biosphere (estimated to be around 10^{31}, about ten times more abundant than bacteria (Breitbart and Rohwer, 2005; Suttle, 2005; Edwards and Rohwer, 2005; Clokie et al., 2011)) and only a minuscule fraction has been identified (Mokili et al., 2012). Diverse phenomena critical to the biology of microbes have been described to be driven by viruses, especially in response to rapid environmental change (Suttle, 2005). Therefore, the view that viruses are "only" parasites is no longer valid. In the environment, viruses are able to transfer and store genetic information of their host population and influence entire biogeochemical cycles. Hence, some viruses are pathogens causing important diseases (in humans, animals or plants) but the great majority can play important roles in regulating entire ecosystems.

The field of "virology" also needs to deal with a variety of different

---

* Corresponding author at: European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany.
  E-mail address: manja@uni-jena.de (M. Marz).
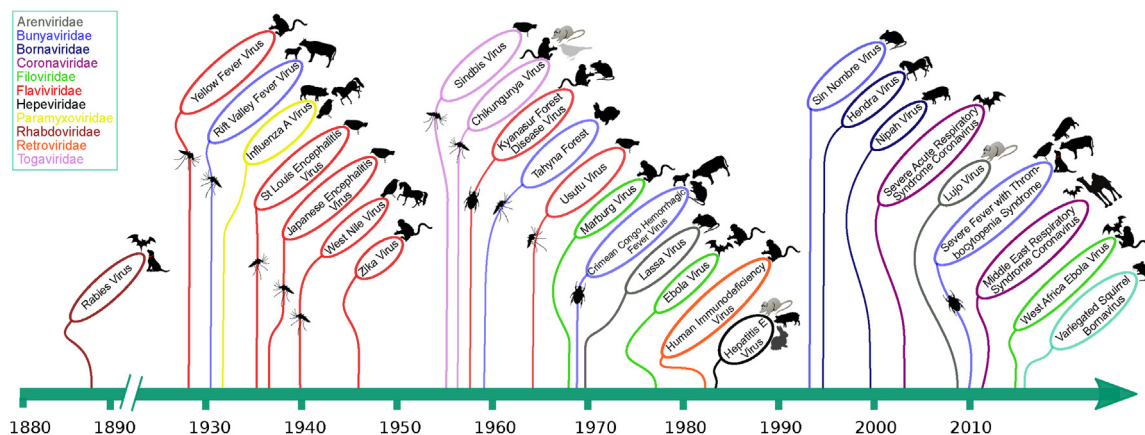 [1] http://www.who.int/csr/research-and-development/r_d_blueprint_plan_of_action.pdf.

Fig. 1. Unknown/new viruses emerge all the time. Figure is an extended and redrawn version of https://www.microbiologysociety.org/publication/past-issues/zoonotic-diseases.html.

viruses with fundamentally different biological properties including their genetic organization, replication strategies, host range and host interactions (commensal, antagonists, mutualistic). Importantly, viruses evolve very rapidly and can quickly vary their genomes in response to various selective pressures including the most sophisticated control measures deployed by host (eukaryotic and prokaryotic) immune systems and/or therapeutic interventions.

There are many fundamental question in virology that need to be tackled. For example, how can we capture the full diversity of virus families in different hosts and environments? How do viruses evolve and how important is recombination in viral evolution? Is there a single common viral origin or do we find clearly independent origins? How can we identify the dynamic gene pool carried by viruses in various ecosystems? Many other questions will help us to develop strategies to control and treat viral diseases but also to understand the broader ecological role of viruses.

The power of new genome sequencing technologies, associated with new tools to handle "big data", provide unprecedented opportunities to address fundamental questions in virology. We would like to emphasize that many of the common questions raised in virology require specific bioinformatics support (Chang, 2015; Marz et al., 2014) and require the combined expertise of both bioinformaticians and virologist. This is because highly specific computational approaches are becoming absolutely necessary to address some of the key questions that we highlight here, in addition to the many other questions being addressed in virology laboratories across the globe. Approaches to tackle these important questions are discussed elsewhere (Marz et al., 2014).

2. Is bioinformatics ready to go viral?

There have been remarkably few bioinformatics communities focusing on viruses. With few exceptions, viral genomes are therefore rather poorly annotated and few computational tools and techniques

have been developed specifically to analyze the idiosyncratic features of individual virus families.

Technically, the small size of viral genomes makes it possible to sequence large numbers of isolates, usually in clinical contexts, an advantage that is generally unavailable for any other living system. This flood of sequencing data in itself calls for specific methods of analysis, which so far are partially available at best. Nevertheless, the current sequencing technologies available for viral genomes pose challenges because most analysis steps are not easily automated and every method approach has its own peculiar set of technical limitations (Marz et al., 2014). However, by integrating bioinformatic methods, it could in future be possible to predict viral evolution in patients just based on individual virus population characteristics, such as whether an individual contains a low prevalence virus population with limited genetic variation. Here, the ultimate goal would be to forecast the course of a virus infection and to adjust therapeutic treatments accordingly.

Clearly there is an emerging need for an integrated workflow combining the different processing steps in viral diversity studies (Hufsky et al., 2018). Such a workflow could then assist clinicians and virologists on a daily basis to discover and characterize the underlying virus populations that are causing disease. Attempts have recently been made towards achieving this aim, some of which are listed in the following section and Table 1.

3. Virus related databases and tools

A major challenge for algorithm and software development in the big data field is the biodiversity of viruses with its coverage of multiple scales and its high complexity (Hölzer and Marz, 2017). Recently, a handful of new databases and tools have become available to virologists that will be discussed in the following section. A summary of some of these databases and tools is shown in the first column of Table 1.

Table 1
List of *selected* virus bioinformatical databases and tools. Further specific details can be found at http://evbc.uni-jena.de/tools.

Databases	De novo assembly	Secondary structure	Sequencing and annotation	Phylogenetic inference
DIGS (Database-integrated, 2017)	AV454 (Henn et al., 2012)	mfold (Zuker, 2003)	ATHLATES (Athlates, 2017)	AdaPatch (Adapatch, 2017a)
EpiFlu (Shu and McCauley, 2017)	SPAdes (Nurk et al., 2013)	LocARNA (Will et al., 2007)	GLUE (Glue, 2017)	AntiPatch (Antipatch, 2017b)
HCV (Kuiken et al., 2005)	RIEMS (Scheuch et al., 2015)	LRIscan (Fricke and Marz, 2016)	PriSM (Prism, 2017)	AntigenicTree (Antigenictree, 2017)
HIV (Druce et al., 2016)	V-FAT (Charlebois et al., 2017)	RNAalifold (Hofacker, 2007)	Tanoti (Tanoti, 2017)	
ICTV (The international, 2017)	VICUNA (Yang et al., 2012)	RNAfold (Gruber et al., 2008)		
ViPR (Pickett et al., 2012)	VrAP (Fricke et al., 2017)			
ViralZone (Hulo et al., 2011)	SOAP (Luo et al., 2012)			
VVR (Hatcher et al., 2017)				