

Interface Between Bioinformatics and Virology

Martin Hölzer

Friedrich Schiller University, Germany
Email: martin.hoelzer@uni-jena.de

Abstract

In the last years, we have witnessed both the emergence of new viral diseases (e.g. MERS, SARS) and therefore the re-emergence of known diseases in new geographical areas (e.g. Zika, Dengue and Chikungunya). Virologists have traditionally concentrated on studying viruses that cause disease in humans, animals or plants. However, there has been estimated around 1031 viruses within the biosphere and only a minuscule fraction has been identified, yet. On the other hand, 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 and for the need to bring together the expertise of bioinformaticians and virologists. For this mission, the ECVBC (European Virus Bioinformatics Center) was founded on 8th March 2017 in Jena, Germany. The ECVBC has about 100 founding members from over 50 research institutions distributed across 13 European countries. I will present first attempts to tackle with virus-specific programs open questions in (co-)phylogeny, high-throughput sequencing data analysis, virus detection, virus-host interaction and host barriers.

Viruses are the cause of a considerable burden to human and animal health. In recent years, we have

witnessed both the emergence of new viral diseases, and the re-emergence of known diseases in new geographical areas. The increased global risk of viral emergence is thanks to a spread of social, environmental and ecological factors. Climate change, deforestation, urbanization, and therefore the unprecedented mobility of products, 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. Diseases of livestock affect food security and inflict considerable economic damage. Diverse phenomena critical to the biology of microbes have been described to be driven by viruses, especially in response to rapid environmental change. Therefore, the view that viruses are “only” parasites is no longer valid.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 viruses with fundamentally different biological properties including their genetic organization, replication strategies, host range and host interactions. 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 questions in virology that need to be tackled. For example, how can we capture the complete diversity of virus families in several hosts and environments? How do viruses evolve and the way important is recombination in viral evolution? Is

there one common viral origin or can 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.

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 are developed specifically to research 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. However, by integrating bioinformatic methods, it could in future be possible to predict viral evolution in patients just supported individual virus population characteristics, like whether a private 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. Such a workflow could then assist clinicians and virologists on a day to day to get 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.

Scientists have united during a common search to sequence, store, and analyze genes and proteins. Understanding what has been done and what we can do in silico is essential in designing new experiments. Taking into consideration the role-play of viruses in causing infections and their use in biotechnology, this review describes the appliance of protein bioinformatics in virology. Therefore, a selection of important features of viral proteins like epitope prediction, protein docking, subcellular localization, viral protease cleavage sites, and computer-based comparison of their aspects are discussed. This paper also describes several tools, principally developed for viral bioinformatics. Prediction of viral protein features and learning the advances during this field can help a basic understanding of the connection between an epidemic and its host. The future of virus bioinformatics depends on rapid specific bioinformatical software development, the establishment of useful virus-specific databases and tools, and thus the establishment of joint interdisciplinary research projects. It also requires immediate actions, including graduate summer schools, ring trials, courses for principal investigators, and annual meetings and workshops. These cannot be met by individual countries acting alone. The ecu Virus Bioinformatics Center (EVBC), consisting of virologists and bioinformaticians from everywhere Europe, was recently founded to coordinate efforts during a new era of virus bioinformatics. The EVBC hopes to fill a number of elemental outstanding knowledge gaps facing virus research.