frontiers in Virology

SPECIALTY GRAND CHALLENGE
published: 22 April 2021
doi: 10.3389/fviro.2021.684608

Bioinformatics Algorithms and Predictive Models: The Grand Challenge in Computational Virology

Manuela Sironi[1]* *and Lars Kaderali*[2]*

[1] Computational Biology Unit, Scientific Institute IRCCS E. Medea, Bosisio Parini, Italy, [2] Institute of Bioinformatics, University Medicine Greifswald, Greifswald, Germany

Keywords: phylodynamics, multi-omic, metagenomics, metaviromics, molecular epidemiology

While we are writing these lines, the first year of the COVID-19 pandemic is coming to an end. The worldwide emergency is not over, though, and SARS-CoV-2 is still circulating widely in most countries. Together with its disruptive effects on health systems, societies, and economies, the pandemic has also triggered an unprecedented international effort by the scientific community. Research has proceeded at enormous speed. An immense amount of data has been generated and almost immediately released for public use. The spread of SARS-CoV-2 has been followed in real-time by active sequencing and application of phylodynamic approaches. During the first pandemic year, more than 400,000 complete viral genomes have been deposited in repositories such as GISAID (https://www.gisaid.org/) and ViPR (https://www.viprbrc.org/). Making sense of such a wealth of sequence data is being a challenge *per se*, only partially met by the existence of pipelines for phylodynamic analysis such as Nextstrain (1).

Never in the past has the relevance of bioinformatic and predictive tools been more central in the field of virology as today. SARS-CoV-2 has brought along a huge health burden, but also a deeper awareness that scientific progress can no longer be effective without extensive systems for data storage, sharing and analysis, as well as computational tools dedicated to molecular epidemiology, NGS data analysis, prediction of drug targets, multi-OMIC data integration, and many other applications.

The birth of bioinformatics is often placed in the year 1962, when, on punch-cards, Margaret Dayhoff and Robert Ledley developed COMPROTEIN, a FORTRAN-based program to determine protein sequences from peptide sequencing data obtained by the Edman degradation method (2, 3). Interestingly, one of the first problems that computational tools were designed to address was the assembly of hundreds of short peptide sequences into a whole protein sequence. With due differences, this is not so distinct from one of the major applications of present-day bioinformatics algorithms, namely the reconstruction of genomes, meta-genomes, and transcriptomes from millions of short sequence reads.

In the years following the development of COMPROTEIN, the advancement of computational tools closely paralleled that of molecular and cellular biology methodologies, with a major breakthrough occurring when nucleic acid sequencing methods became available. The first complete genomes to be obtained were those of two viruses, specifically of bacteriophages PhiX174 (4) and MS2 (5), in 1976–1977. Since then, technological progress has allowed a constant increase in the number of available sequence data, with the rise becoming exponential from 2001 onward, thanks to the advent of NGS and other high-throughput technologies. Going back to human coronaviruses as a test case, fewer than 100 complete viral genomes were obtained during the SARS-CoV epidemic of 2002–2003, and around 1,000 MERS-CoV complete sequences were generated a few years later, since the first case was registered in 2012 (6). These numbers are in striking contrast with the more than 400,000 SARS-CoV-2 genomes deposited in public databases in the past year.

OPEN ACCESS

Edited and reviewed by:
Akio Adachi,
Kansai Medical University, Japan

***Correspondence:**
Lars Kaderali
lars.kaderali@uni-greifswald.de
Manuela Sironi
manuela.sironi@lanostrafamiglia.it

Specialty section:
This article was submitted to Bioinformatic and Predictive Virology, a section of the journal Frontiers in Virology

Received: *23 March 2021*
Accepted: *30 March 2021*
Published: *22 April 2021*

Citation:
Sironi M and Kaderali L (2021) Bioinformatics Algorithms and Predictive Models: The Grand Challenge in Computational Virology. Front. Virol. 1:684608. doi: 10.3389/fviro.2021.684608

These figures very well-represent the general trend in all fields of research, not only virology and not only nucleic acid sequencing–the generation of huge amounts of data. Extracting biological and clinical knowledge from such data using analytical and predictive computational tools is the overarching grand challenge of computational biology.

Metagenomics and metaviromics approaches have revealed that viruses are the most abundant and most genetically diverse entities in the biosphere [reviewed in Koonin et al. (7)]. The amount of new viral genetic data that are generated through metagenomics has revolutionized the field of virology to such an extent that the International Committee for Taxonomy of Viruses (ICTV) has made the decision to classify new viral species (or higher taxa) solely on the basis of metagenomic data (8). The ICTV has also recently approved the establishment of high taxonomic ranks (e.g., order, realm, kingdom, phylum) to facilitate virus classification (7, 9, 10). Still, major challenges remain in the classification of millions of viruses, as members of the ICTV have recently highlighted (9). Moreover, whereas genome sequences are readily generated, little is known about the biological, evolutionary and ecological characteristics of most newly discovered viruses. It is likewise largely unknown whether and which viruses represent potential threats for humans, other animals or plants/crops. On the one hand, the SARS-CoV-2 pandemic has clearly shown that we have remarkably little capacity to predict which viruses are likely to spillover to humans and even less ability to predict their phenotype in terms of host range, virulence and transmissibility. On the other hand, despite the huge amount of virus genome data that are being generated worldwide, we still have little clues as to where some of the most widespread human pathogens (e.g., HCV) came from and when or how this happened.

In recent years, enormous progress has been made in the development of computational tools to study virus evolution and track viral spread in time and space. The SARS-CoV-2 epidemic is not the only example of real-time phylodynamic analysis. Similar, although smaller scale, approaches have been applied to study the cross-country Ebola virus epidemic of 2013-2016, the Zika virus pandemic, as well as the surge in Lassa virus outbreaks in West Africa (11–17). We have learned a lot, but we also lack major insights into several pivotal issues. First and foremost, the role (or lack thereof) of viral genetic diversity in disease presentation. Many challenges lie ahead in data analysis and interpretation in the field of virology–and this is the topic of the bioinformatics and predictive virology section of Frontiers in Virology: Bioinformatics algorithms and predictive models to understand viral evolution and phatogenicity, virus-host interactions, and linking viral (and host) diversity with manifestation and presentation of disease. We are looking forward to working with all of you on this exciting topic and pursuing this challenge.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

1. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. (2018) 3423:4121–3. doi: 10.1093/bioinformatics/bty407
2. Dayhoff MO, Ledley SL. Comprotein: a computer program to aid primary protein structure determination. In *AFIPS '62 (Fall): Proceedings of the December 4-6, 1962, Fall Joint Computer Conference*. New York, NY (1962) 262–274. doi: 10.1145/1461518.1461546
3. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. *Brief Bioinform*. (2019) 206:1981–96. doi: 10.1093/bib/bby063
4. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. (1977) 2655596:687–95. doi: 10.1038/265687a0
5. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*. (1976) 2605551:500–7. doi: 10.1038/260500a0
6. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med*. (2012) 36719:1814–20. doi: 10.1056/NEJMoa1211721
7. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global organization and proposed megataxonomy of the virus world. *Microbiol Mol Biol Rev*. (2020) 842:e00061–19. doi: 10.1128/MMBR.00061-19
8. Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: virus taxonomy in the age of metagenomics. *Nat Rev Microbiol*. (2017) 153:161–8. doi: 10.1038/nrmicro.2016.177
9. Kuhn JH, Wolf YI, Krupovic M, Zhang YZ, Maes P, Dolja VV, et al. Classify viruses - the gain is worth the pain. *Nature*. (2019) 5667744:318–20. doi: 10.1038/d41586-019-00599-8
10. Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, et al. Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). *Arch Virol*. (2019) 1643:943–6. doi: 10.1007/s00705-018-04136-2
11. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet*. (2018) 191:9–20. doi: 10.1038/nrg.2017.88
12. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*. (2014) 3456202:1369–72. doi: 10.1126/science.1259657
13. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SL, Andersen KG, et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell*. (2015) 1617:1516–26. doi: 10.1016/j.cell.2015.06.007
14. Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, et al. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science*. (2019) 3636422:74–7. doi: 10.1126/science.aau9343
15. Ehichioya DU, Dellicour S, Pahlmann M, Rieger T, Oestereich L, Becker-Ziaja B, et al. Phylogeography of lassa virus in Nigeria. *J Virol*. (2019) 9321:e00929–19. doi: 10.1128/JVI.00929-19
16. Faria NR, Azevedo RD, Kraemer MU, Souza R, Cunha MS, Hill SC, et al. Zika virus in the Americas: early epidemiological and genetic findings. *Science*. (2016) 352:345–9. doi: 10.1126/science.aaf5036
17. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, et al. Zika virus evolution and spread in the

Americas. *Nature*. (2017) 5467658:411–5. doi: 10.1038/nature22402

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.