BioHackrXiv

SB4ER: an ELIXIR Service Bundle for Epidemic Response

Castrense Savojardo[1], Pier Luigi Martelli[1], Giulia Babbi[1], Marco Anteghini[2], Matteo Manfredi[1], Giovanni Madeo[1], Emidio Capriotti[1], Jumamurat R. Bayjanov[3], Margherita Mutarelli[4], and Rita Casadio[1]

1 Biocomputing Group, Depertment of Pharmacy and Biotechnology, University of Bologna, Italy **2** LifeGlimmer GmbH, Berlin, Germany **3** Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands **4** Telethon Institute of Genetics and Medicine, Pozzuoli, Naples, Italy

BioHackathon series:
COVID-19 BioHackathon
Virtual conference 2020

Submitted: 29 Jan 2021

Introduction

Epidemic events, both affecting humans or livestock and plants, have occurred throughout all human history, often conditioning or even overturning it. To understand this concept, it is enough to know that the term "epidemic" appeared for the first time in Homer and was the title of a Hippocratic treaty (Martin & Martin-Granel, 2006). Nevertheless, epidemics are of obvious relevance also in today's times. Examples for this are the spreads of new pathogenic viruses, which we observed in the last few decades. This included both mutations of the classical influenza viruses, e.g. the influenza A (H1N1) virus (Sullivan, Jacobson, Dowdle, & Poland, 2010) and viruses that crossed the animal-human divide (Mostafa, Abdelwhab, Mettenleiter, & Pleschka, 2018). These spreads of new pathogens, for which there is no pre-existing immunity or an available vaccine, may threaten the onset of an epidemic, or even a pandemic, event.

Given this context, the ELIXIR (https://elixir-europe.org/) Service Bundle for Epidemic Response (SB4ER), which we developed during the BioHackathon Europe 2020 (https://www.biohackathon-europe.org/), aims to provide tools and resources to collect and analyse data on new pathogens (bacteria and viruses) and their relation to hosts (humans, animals, plants). SB4ER collects more than 140 resources endowed with comprehensive annotations, which allow the user to select the most appropriate tools and databases. Moreover, resources are organized in concept maps according to five possible use scenarios, which formally defines scientific needs in the context of the study of a new pathogen.

Service Bundles (https://elixir-europe.org/services/service-bundles) aim to group tools, services, people and training materials devoted to particular use scenarios. These resources are organized providing information about their interoperability and applicability to specific use cases. Then, a Service Bundle simplifies the user's access to resources available to work in a given research field. Moreover, Service Bundles helps to highlight the presence of gaps to be filled in a given research area.

Use Scenarios Definition and Resources Collection

As a first step, we broke down the most common operations that need to be performed when analyzing a new pathogen into five use scenarios.

The first scenario corresponds to the primary analysis of a new genome sequence. It begins with the assembly of the genome starting from NGS data and it has the final goal of performing functional annotation of both the genes and the coded proteins. At this step, we begin to understand how this pathogen will behave and we produce important data that later on will be used as input, i.e. an assembled reference genome and a gene to protein mapping.

For the second use scenario, we focused on the study of the phylogenetic history of the new pathogen. We included here all of the steps needed to perform multiple sequence alignments and to obtain phylogenetic trees linking the new genome to known species. This is a fundamental step, as it allows us to transfer the knowledge acquired in previous scientific efforts to the new problem we are facing.

If the goal of the second scenario is to understand the relation between the new pathogen and other species, in the third we focused instead on the variations between all of the sampled individuals. Starting from the comparison of different genomes, at the end of the work described in this case, we should have a clearer understanding of the impact of mutations occurring in the pathogen's population.

The three use scenarios described until now are the most general, including and summarizing the basic steps performed when studying a new genome; their end results are the starting point for further analysis. For this reason, we specified two more specific scenarios, where we grouped the operations that need to be performed respectively for assessing the antibiotic/antiviral resistance from the genome sequence and for discovering important factors for vaccine design.

We analyzed all the use scenarios described above in terms of operations needed to perform them, producing a flow chart for each one of them. Once the use scenarios and the relative steps were defined, we went on to resources collection. For each one of the operations identified, we found available tools and pipelines designed to perform them and databases providing useful information on the problem at hand. The collected resources were, in most cases, present in bio.tools (Ison et al., 2016). However, each resource has been endowed with annotations describing it in details. The complete list is made available along with annotations in order to allow aware choices of the resources to use and to provide useful information about their employment. Annotations provided include: a brief description of the resource; the year of first publication and eventual update; the reference; the specific use scenario and operation for which the resource may be used; information about whether or not the resource is specialized on virus or bacteria; information about the distribution and the availability of the resource (web server, standalone version, license…); a link to the respective bio.tools page, if any; technical aspects, i.e. input and output formats, main dependencies (if present). Table 1 reports some data relative to the collected resources.

Table 1

Map	N. of Resources	Included in bio.tools	Databases
Map 1	38	35	9
Map 2	27	25	4
Map 3	28	25	3
Map 4	17	13	8
Map 5	33	31	4
Total	143	129	28

Concept maps

As a final step for our work, each one of the defined use scenarios and relative resources associated was organized in a concept map, via the service provided by the Training eSupport System (TeSS) (https://tess.elixir-europe.org/). These concept maps are visual, step-by-step protocols describing all the operations, represented by a node, required to perform a given task. In our case, each node was endowed with a list of available resources, useful for performing the task at hand.

Map 1: Primary Analysis of Genome Sequence

The first map describes the basic steps for genome assembly starting from NGS data, gene prediction and functional/structural annotation of encoded proteins. Thus, starting from the raw nucleic acid sequence, the concept maps firstly introduces quality check and trimming, then genome assembly, to obtain a genome report. The genome obtained may be deposited in specialized databases. In this perspective, tools useful to prepare data for this purpose are reported.

Following the proposed workflow, the genome is then processed to obtain single genes using tools for gene prediction. Once obtained, this information may also be shared via database deposition. Moreover, genes are functionally annotated and associated to a protein sequence through ID mapping in databases or DNA/RNA translation tools. Finally, proteins are also functionally annotated and studied in terms of protein-protein interaction using specific tools and databases.

Map 2: Multiple Sequence Alignment and Phylogeny

In the second map, tools and resources are organized to compare genes and genomes and to reconstruct phylogenetic history for the new pathogen. The first step performed in this task is an orthology/core genome analysis, in which, having as input genes and a list of genomes, core genes are identified. Core genes are defined as genes present in all the strains of a species. These genes are then used to build multiple sequence alignments, from which super alignments are obtained via filtering with specific tools. From super alignments, phylogenetic trees are reconstructed. Moreover, as a final step, tools for phylogenetic trees visualization are proposed.

Map 3: Mutation Variant Types

The starting point of the third use scenario could either be raw NGS data obtained from sequencing or assembled genomes. In both cases, if we want to extract variation data, we need to analyze different sequences coming from the sequencing of many individuals. Those could come from the same laboratory or be taken from a common repository where researchers from around the globe are storing the results of their experiments.

From this collection of sequences, Genome Comparison and Variant Calling can be performed with different tools in order to extract a list of all of the observed variations mapped onto a common reference genome. This information is usually stored into a file format called VCF (Variant Call Format).

From here two main analyses can be performed. The first consists of retracing the origin of the mutations, performing an intra-species phylogenetic analysis. The second is to classify the different variants that we identified, and for non-synonymous variations occurring on protein-coding regions of the genome, we can further study their effect on the stability and the function of the coded protein.

Map 4: Assessing Antibiotic/Antiviral Resistance

The fourth map addresses the problem of assessing antibiotic or antiviral resistance from the genome sequence, highlighting the presence of genes or variations related to these phenotypes. This is realized having as starting point nucleic acid sequencing data and genes/genome, possibly derived via the application of the first map. All these input data are used to filter out antimicrobial resistance-related genes, using the information obtained from ad hoc databases. These selected genes are analyzed in terms of gene location (plasmid or chromosome) and for the presence of mutations, both known or novel, for which the possible impact may be predicted. All these information are finally aggregated to obtain a report for antimicrobial resistance referencing terms from the Antibiotic Resistance Ontology (ARO).

Map 5: Important Factors for Vaccine Design

The preliminary work on vaccine design that is described in this use scenario starts from two main sources of data. The first is the assembled and annotated genome that comes directly from the output of the first concept map, alongside a list of coded protein sequences. The second is a structure database from where we can extract proteins homologous to the ones we are studying so that we are able to build 3-dimensional models of their structure. Using that information we can then detect exposed proteins and identify epitopes that can be used in the development of new vaccines for the pathogen. Starting from the sequences alone, another step that different tools can perform in an automated way is the preliminary DNA vaccine design.

Discussion

During the Biohackathon Europe 2020, we completed the release of a Service Bundle for Epidemic Response. We were able to identify 5 different use scenarios representing challenges that biologists may face when studying a new pathogen. We then translated them into concept maps that will guide them while providing a list of useful databases and tools for performing all of the different tasks in an automated way. We think that SB4ER may be a powerful tool in the hands of researchers for performing preliminary and immediate analysis of newly discovered pathogens. The step-by-step protocols, endowed with extensive and complete annotations will speed up the analysis, allowing quick access to relevant information and helping in designing an initial response to possible epidemic events.

Future work

Now that the Service Bundle is released, there is still much work to be done for improving it and keeping it up to date. The main focus should be put on expanding the use scenarios that we defined. New specific cases can be defined, covering broader fields of research that share the goal of providing a strong and quick response against new pathogens. At the same time, the generic scenarios will need to cover more operations in order to help to retrieve all the data needed by the specific ones.

Another important effort that we will need to make is to add new fields of information to the spreadsheet containing the description of the tools and databases we collected. This will require both extensive testing of the methods and in-depth reading of the literature in order to provide to the users clear and accessible meta-data on the available resources.

At the same time, we will periodically update the list with the new state-of-the-art methods that researchers will develop in the forthcoming years. For this reason, we will gladly accept any report of methods that the scientific community thinks should be included, making it more visible and easy to be used by researchers working in this field.

GitHub repositories and data repositories

- https://sb4er.github.io/index.html

Acknowledgements

This work was done during the BioHackathon Europe 2020 Online, organized and funded by the ELIXIR Hub in November 2020. We thank the organizers for the opportunity.

References

Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., Anthon, C., et al. (2016). Tools and data services registry: A community effort to document bioinformatics resources. *Nucleic Acids Research*, *44*(D1), D38–D47. doi:10.1093/nar/gkv1116

Martin, P. M., & Martin-Granel, E. (2006). 2,500-year Evolution of the Term Epidemic. *Emerging Infectious Diseases*, *12*(6), 976–980. doi:10.3201/eid1206.051263

Mostafa, A., Abdelwhab, E. M., Mettenleiter, T. C., & Pleschka, S. (2018). Zoonotic Potential of Influenza A Viruses: A Comprehensive Overview. *Viruses*, *10*(9), 497. doi:10.3390/v10090497

Sullivan, S. J., Jacobson, R. M., Dowdle, W. R., & Poland, G. A. (2010). 2009 H1N1 influenza. *Mayo Clinic Proceedings*, *85*(1), 64–76. doi:10.4065/mcp.2009.0588