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

Published by BioHackrXiv.org

Introduction

Epidemic events, both affecting humans, livestock or 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.

An early response is, in all cases, essential to limit the spread and the consequent damages of any epidemic event. Computational analysis of genomes can substantially help the elucidation of biological mechanisms at the basis of the infection, of disease insurgence and can facilitate the discovery of molecular targets for drugs and vaccines.

Tools, data, workflows for the study of viral and bacterial genomes have been implemented during the years but they are mostly spread in different locations and they still lacks an organization into conceptual maps implementable as bioinformatics workflows. Many of these resources are available within the ELIXIR infrastructure.

ELIXIR Service Bundles (https://elixir-europe.org/services/service-bundles) aim at grouping 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. The ultimate goal of a Service Bundle if to simplify the access to resources available for a specific research field and to help end users in the proper selection of tools with respect to a specific scientific need. Moreover, Service Bundles can be used to identify potential gaps to be filled in a given research area.

Here, we describe the outcomes of a project we carried out during the BioHackathon Europe 2020 virtual event. The project aimed to build a Service Bundle for Epidemic Response (SB4ER), namely a collection of tools and resources for the genomic analysis of new pathogens (either bacteria and viruses) and their relation to hosts (humans, animals, plants). Mainly focusing on the portfolio of ELIXIR resources (as available on bio.tools), the SB4ER collects more than 140 tools endowed with comprehensive annotations, allowing users 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.

SB4ER: Use Scenarios Definition and Resources Collection

In the first part of the project, we focused on the definition of common use scenarios arinsing during the genomic analysis of new pathogens. Three general scenarios were identified and definied, corresponding to basic tasks that are preliminary to all subsequent analyses:

- Primary analysis of genomic sequence of new pathogen.
- Phylogenetic analysis.
- Analysis and characterization of varitions

We then defined two more specific scenarios, corresponding to advanced analytical steps:

- Assessing the antibiotic/antiviral resistance.
- Vaccine and drug design.

We analyzed all the use scenarios described above in terms of operations or steps needed to perform them, producing a flow chart for each one. We then started collecting tools and resources that are relevant for each use scenario. Specifically, for each step, we mapped available tools and pipelines as well as key databases. The vast majority of collected resources were already present in bio.tools (Ison et al., 2016). Additional information were also collected for selected resources, including:

- A brief description of the resource.
- The Relevant literature linked to the resource.
- The link to specific use scenario and operation for which the resource may be used for.
- Information about the distribution and the availability of the resource (web server, standalone version, license).
- Technical aspects, e.g. input and output formats, main dependencies (if present).

Table 1 reports a summary of the collected resources.

Table 1: Resources included in SB4ER.

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

Concept maps

Use scenarios and relative resources associated were organized in concept maps, describing how the different analytical steps are interconnected in the context of a specific use scenario. Concept maps were designed using the dedicated formalism implemented by the ELIXIR Training eSupport System (TeSS) (https://tess.elixir-europe.org/) platform. These concept maps are visual, step-by-step protocols describing all the operations, represented by a node, required to perform a given task. Each node is associated with a set of tools and/or resources available for performing the task at hand.

Use Scenario 1: Primary Analysis of Genome Sequence

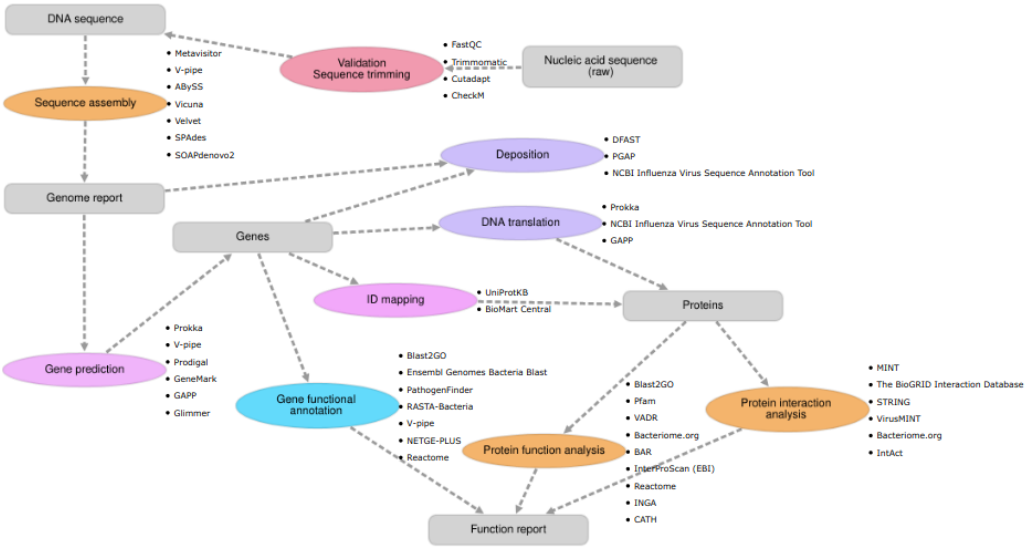

Figure 1: Concept map for primary analysis of pathogen genome sequence.

The first concept map (Figure 1) 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 include canonical pre-processing steps such as 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.

Use Scenario 2: Multiple Sequence Alignment and Phylogeny

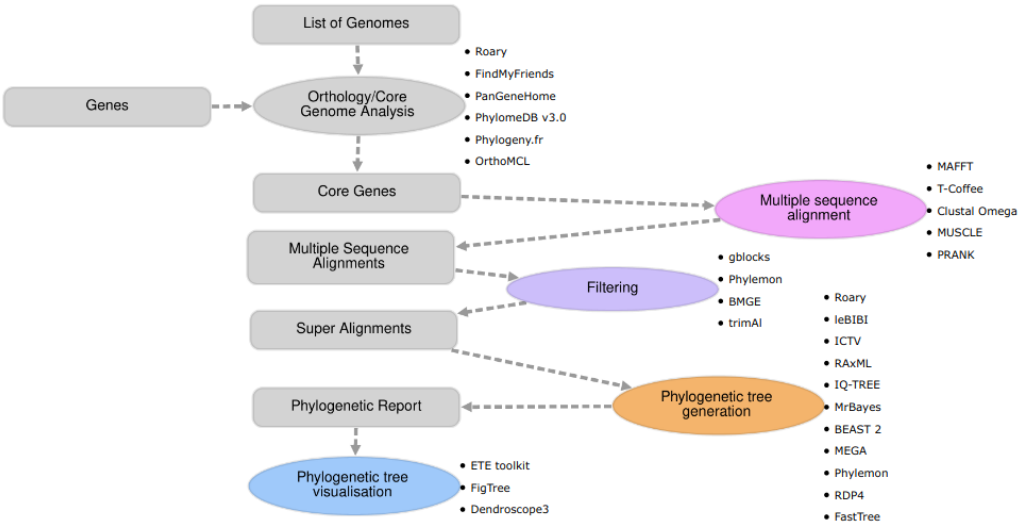

Figure 2: Concept map for phylogenetic analysis.

In the second map (Figure 2), tools and resources are organized to compare genes and genomes and to reconstruct phylogenetic history for the new pathogen, performing a canonical phylogentic analysis. 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.

Use Scenario 3: Analysis of variants

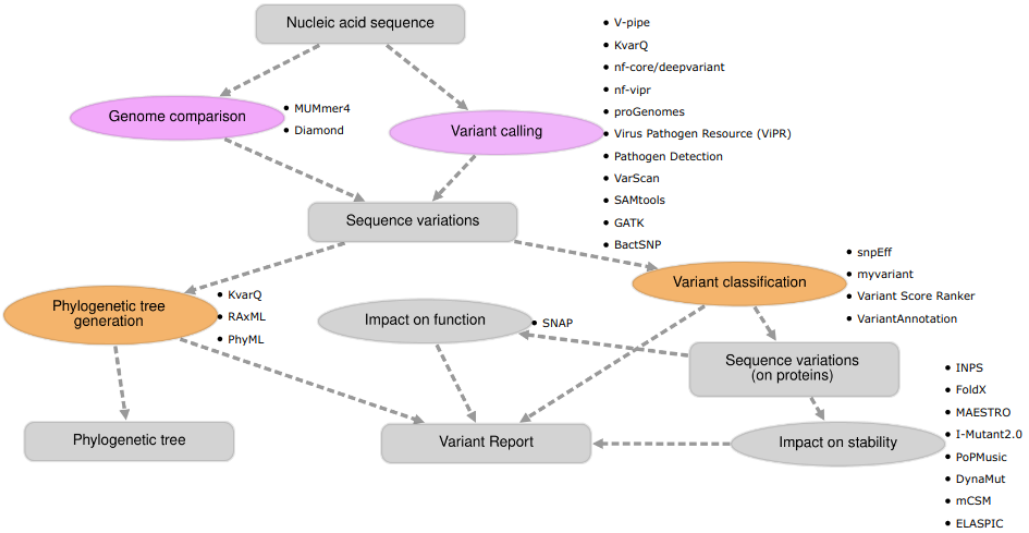

Figure 3: Concept map for analysis of variants.

The starting point of the third use scenario (Figure 3) 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 (potentially retrieved from public databases).

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.

Use Scenario 4: Assessing Antibiotic/Antiviral Resistance

Figure 4: Concept map for antimicrobial/antiviral resistance.

The fourth map (Figure 4) 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).

Use Scenario 5: Vaccine/drug design

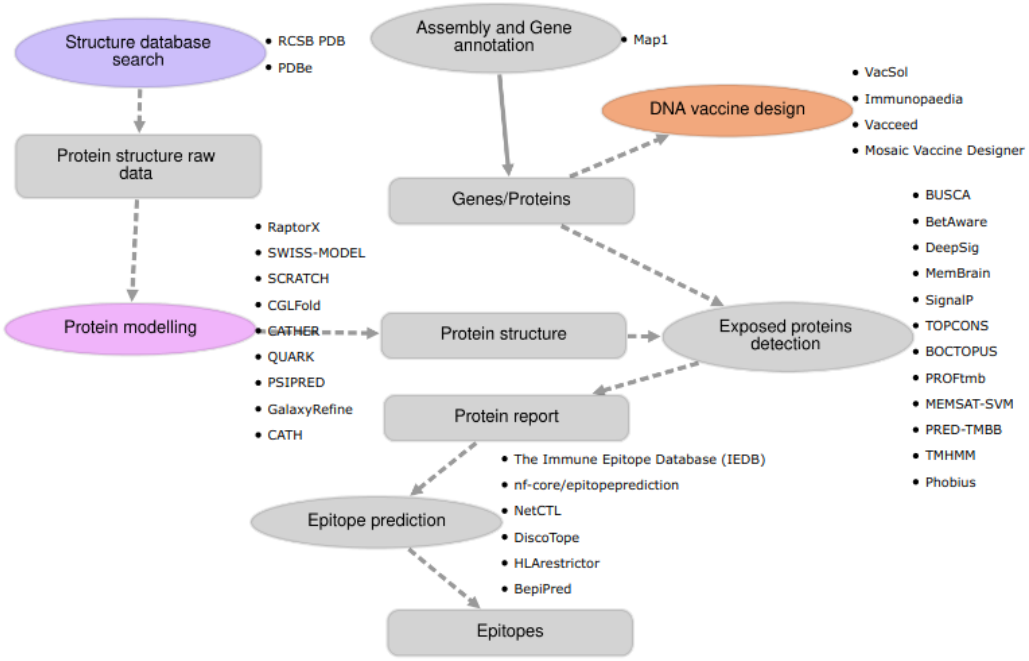

Figure 5: Concept map for vaccine/drug design.

The preliminary work on vaccine/drug design (Figure 5) that is described in this use scenario starts from two main sources of data: i) the assembled and annotated genome, alongside a list of coded protein sequences and ii) a structure database (e.g. PDB) that can be used for comparative protein modelling. Exposed proteins can be then detected and epitopes identified. The latter can be used in the development of new drugs and vaccines for the pathogen.

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 for performing preliminary and early 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

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.

Efforts will be also devoted to periodically update the list of tools and resources, to be aligned with newl releases.

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