During the last nine months, we have been used to following the fight against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the resulting coronavirus disease (Covid-19) daily through news and social media. We hear about new mutations and what they might mean for the disease. We also hear about possible breakthroughs in vaccine development, and in finding new drugs that may help in the fight against the virus. It is easy for us to picture clinicians and experimental biologists at work in hospitals and at lab benches trying to understand how the virus works, and how the disease hits. It is maybe harder for us to picture computer scientists at work as equally important players in this quest, but in reality, most of the findings and breakthroughs that we rely on in fighting Covid-19 (and other diseases) depend on the availability of numerous computational tools and other information about the patient that may explain why they are badly affected, such as age, gender, and clinical data such as blood pressure, other diseases etc, making it possible to see the whole picture. And the data should be represented using standardized vocabularies in a machine-readable form, enabling integration and further processing in other contexts.

Databases containing all of the building blocks of SARS-CoV-2

In order to identify the new virus as a SARS-like virus, the genetic information of the virus, i.e. represented by the nucleotide sequence, had to be compared with the sequences of all previously known viruses that had been stored in the nucleotide databases, for example the European Nucleotide Archive (ENA). In its simplest explanation, this happens by sending the nucleotide sequence of the new virus to a tool that searches the database and returns the nucleotide sequences that are most similar to it. The comparison tools that are used will also give a similarity score that tells us how similar the virus is to the various other viruses that have already been stored in the database, and in this case it returned the previously described 2002 SARS virus, or the SARS-CoV to be correct. Using other bioinformatics tools, we can map out which parts of the new virus that are very similar to the previous SARS-CoV, and which parts that are different. We can track the changes in the nucleotide sequence, or the mutations, as the new virus spreads, using variations of the same methods, and we can use yet other bioinformatics tools to predict what these changes might mean for the different virus strains, whether they for example are likely to make them more infectious, or severe, or maybe both.

When it comes to the bioinformatics comparison- and prediction tools, we are quite lucky that these have been openly available and perfected for as long as we have been able to sequence nucleotide sequences, and therefore we did not have to start all over and make brand new tools for this particular situation. The same goes for the databases, we had a head start against the new virus because we immediately were able to see that it was quite similar to its relative, the SARS-CoV which we now after two decades of research know quite a bit.

As it turns out, ‘open’ is a key word in the fight against SARS-CoV-2 and Covid-19, and most, if not all, of the efforts taken with regard to understanding the new virus rely on having previous data and computational methods publicly available. This may seem obvious but is in fact not. Maintaining a database and making sure that it is up and running and that everything inside is correct and up-to-date takes enormous efforts and depends on having highly qualified personnel available that knows exactly how to do this. These people of course want to be paid for the important job that they do, and in addition, there are costs associated with storing all of the data on servers. The easiest way to solve this would be to ask all users of the database to pay a fee to cover the costs associated with sustaining it. But researchers tend to not want to pay for services they do not use. We could quickly get a situation with isolated data-main-specific databases that did not ‘talk’ to each other, where databases that did not currently cover a hot topic, lost users and died out due to lack of funding. On a global scale, relatively few researchers probably would have paid subscription fees for a corona virus-specific database in the long term, and in a world where this was the general funding model of biological databases, it could very well have suffered this fate. Biology is very unpredictable, and although epidemiologists have long warned us about the probable emergence of global pandemics, we were completely unprepared for the causative agent being a coronavirus.
ELIXIR is the home for the most crucial European databases

On this backdrop, ELIXIR, the European research infrastructure for life science data, was established in 2013, to bring together and safeguard life science resources developed and maintained all over Europe. The ELIXIR collection of resources is not limited to databases, but also includes computational tools, training material, cloud storage and access to supercomputers. The goal of ELIXIR was, and still is, to coordinate all of these resources so that it is easier for researchers to find and share data, exchange expertise and agree on best practices. The bioinformatics communities in Norway have collaborated on providing similar services to Norwegian researchers for almost two decades, originally funded by the FUGE programme of the Research council, and this structure was the basis for the formation of the Norwegian Node of ELIXIR in 2013.

Having a database included in the list of ELIXIR resources generally increases the visibility and usage of the resource and gives national funders an extra incentive to keep on funding it. In addition to several tools, ELIXIR Norway currently provides LiceBase, a database including all available genomic information about the Salmon louse, and the Marine Metagenomic Portal, that collects all available genetic information about marine microbes, including a collection of marine viruses. We have recently opened a call to include other Norwegian services and hope to see this list grow in the future.

On top of this collection of services, the ELIXIR community has started a process to identify databases that are absolutely crucial for the everyday research of the majority of European, and also international, life scientists. As a result of this, a number of ELIXIR Core Data Resources have been selected through a careful evaluation process (Box 1) and will receive extra attention from ELIXIR to make sure that these always are updated, are up and running, and remain open for life scientists and industry also in the future. In part inspired by the effort by ELIXIR, a global coalition of funders is being formed these days. One of the main objectives is to identify core data resources on the global level and to make sure they may Accessible, Interoperable and Reusable for others.

Why it is only FAIR to have open data

The current Covid-19 pandemic has spread extremely rapidly throughout the world, and has effectively shown how important it is that researchers and clinicians are able to immediately share the results and knowledge that they have gathered with international colleagues, and similarly that they are able to access what others have produced. When new mutations are observed in local outbreaks, like the recent one observed in Trondheim in October, it is urgent to compare the sequence of this particular strain to all other available sequences, both to understand what the mutations may mean biologically, and also to understand how this strain may have spread through the population.

The databases (and all services) provided by ELIXIR follow a set of principles to ensure that data can be found, interpreted and used together with data from other databases, and with the user’s own experimental data. These principles are called FAIR³² and call for researchers to make their data Findable, Accessible, Interoperable and Reusable. The circumstances around Covid-19 are very different from those of the annual influenza outbreak, because of the high level of urgency, we want to encourage all Norwegian Covid-19 researchers to publish their sequences in an open database, for example ENA, and we are eager to help researchers to do this.

In reality, most biological databases initially emerged as a result of research projects funded by a national or international research agency. Some databases have proved to be so important that they are able to attract public and private funding over many years, whereas others struggle. In addition to funding, the FAIR principles, which are crucial for the everyday research of the majority of European, and also international, life scientists, as a result of this, a number of ELIXIR Core Data Resources have been selected through a careful evaluation process (Box 1) and will receive extra attention from ELIXIR to make sure that these always are updated, are up and running, and remain open for life scientists and industry also in the future. In part inspired by the effort by ELIXIR, a global coalition of funders is being formed these days. One of the main objectives is to identify core data resources on the global level and to make sure they may Accessible, Interoperable and Reusable for others.

Why it is only FAIR to have open data

The current Covid-19 pandemic has spread extremely rapidly throughout the world, and has effectively shown how important it is that researchers and clinicians are able to immediately share the results and knowledge that they have gathered with international colleagues, and similarly that they are able to access what others have produced. When new mutations are observed in local outbreaks, like the recent one observed in Trondheim in October, it is urgent to compare the sequence of this particular strain to all other available sequences, both to understand what the mutations may mean biologically, and also to understand how this strain may have spread through the population.

The databases (and all services) provided by ELIXIR follow a set of principles to ensure that data can be found, interpreted and used together with data from other databases, and with the user’s own experimental data. These principles are called FAIR and call for researchers to make their data Findable, Accessible, Interoperable and Reusable. The circumstances around Covid-19 are very different from those of the annual influenza outbreak, because of the high level of urgency, we want to encourage all Norwegian Covid-19 researchers to publish their sequences in an open database, for example ENA, and we are eager to help researchers to do this.

Human data need to be treated in a special way

In order to understand how the virus attacks us, and why the attacks become more
severe for some more than others; we also need to study patients who have caught the disease. We need to compare the genetic sequence, or genome, of the patients who have only been mildly hit or not affected at all. Since each study is limited in size, it is useful to use data from several studies to get a better picture. Again, the FAIR principles come into play. However, working with human data and human genomes gives a whole new set of challenges that we have to face. Privacy legislations ensure that personal information cannot be openly shared and re-used for other applications than for those purposes for which the consent originally was given. Handling these personal data securely is therefore collaborating with the rest of the bioinformatics world to fight Covid-19 will allow us to respond even more strongly to the next emerging pandemic.

REFERENCES:

1. https://www.ebi.ac.uk/ena/browser/home
2. https://www.gisaid.org
4. https://www.nature.com/articles/sdata201618
5. https://mmp.sfb.uit.no
6. https://elixir.no
7. https://facebook.com/EGASupport
9. https://www.eosc-life.eu
10. https://ecrin.org
13. https://www.nature.com/articles/sdata201618
15. https://www.elixir.no
16. https://covid19dataportal.no
17. https://covid19.sfb.uit.no
18. https://www.eosc-life.eu
19. https://ecrin.org
20. https://ecrin.org