SARS-CoV-2 Through the Lens of Computational Biology: How bioinformatics is playing a key role in the study of the virus and its origins

Samuel Alizon, Frédéric Cazals, Stéphane Guindon, Claire Lemaitre, Tristan Mary-Huard, Anna Niarakis, Mikaël Salson, Celine Scornavacca, Hélène Touzet

SARS-CoV-2 Through the Lens of Computational Biology

How bioinformatics is playing a key role in the study of the virus and its origins

In December 2019, the Chinese Center for Disease Control reported several cases of severe pneumonia that resisted usual treatments in the city of Wuhan. This announcement marked the beginning of the COVID-19 pandemic, which caused more than 80 million infection cases and 1.7 million deaths worldwide in 2020 alone and is still raging. The pandemic has given rise to global public health responses and international research efforts of unprecedented scope and speed. This scientific mobilization has yielded remarkable results, enabling a great deal of knowledge accumulation in just a few months: from the identification of the virus and its main proteins to the analysis of its origin and mechanisms. This basic biological knowledge is mandatory for medical advances.

In this document, one year after the beginning of the spread of the disease, we wish to shed particular light on the contribution of *bioinformatics* in all this work. This discipline, at the crossroads of computer sciences, mathematics, biology, and physics, has taken on inestimable importance in modern biology and medicine. It provides computational models, algorithms, software, and guidelines to help the scientific community handle biological data and accelerate research. The discovery and study of the SARS-CoV-2 coronavirus is an emblematic example of these contributions. Bioinformatics methods have been at the heart of several essential milestones: sequencing the virus genome, analyzing its origin and evolutionary dynamics, modeling interacting biological entities at the structural and network scales, and studying host genetic susceptibility. For several of these topics, research on SARS-CoV-2 could benefit from a wide range of off-the-shelf software packages that rely on well-established algorithms developed by the bioinformatics community over the years. For other topics, the analysis of SARS-CoV-2 pushes the limits of knowledge and invites the community to develop new computational models and methods. This work, as a whole, has made it possible to elucidate the nature and the functioning of the novel pathogen. It has contributed to the fight against COVID-19, even if much remains to be done to fully understand the disease and control the epidemic.

The document is organized as follows. In Section 1, we narrate the story of the discovery of the virus from its sequencing in the early stages of the pandemic in China to the development of mass testing and low-cost sequencing globally. In section 2, we explain the pivotal role of modeling approaches in understanding how the virus functions at the molecular level. Finally, in Section 3, we come back to the emergence of SARS-CoV-2, its origin, and its evolution in time and space.

1 Tracking the virus's emergence and progression

1.1 Discovery of a new pathogen

Identifying the causative agent of an unknown infection usually requires the sequencing of its genetic material, which means determining the sequence of nucleotides (As, Ts, Cs, and Gs) that make up its genome. Since the mid-2010s, high throughput sequencing coupled with bioinformatics has made it possible to characterize and analyze an emergent virus's genome in a few days for a few hundred euros. Thus, from December 2019 to January 2020, several hospitals in Wuhan confronted with the disease independently embarked on the sequencing of the unknown pathogen [6, 7, 8, 9, 10]. They all followed approximately the same protocol.

Sequencing of the genetic material. The starting point consists of collecting lung fluids from patients, and then extracting the genetic molecules contained in the sample. The result is a pulmonary microbiome that is ready for sequencing. Genomic sequences cannot be generated outright because the sequencing method can only generate short stretches of sequences, measuring approximately 200 nucleotides at a time. So, after the sequencing step, the raw data is a soup of hundreds of thousands of short nucleic sequences, called *reads*, which are intended to randomly cover the initial genomes.

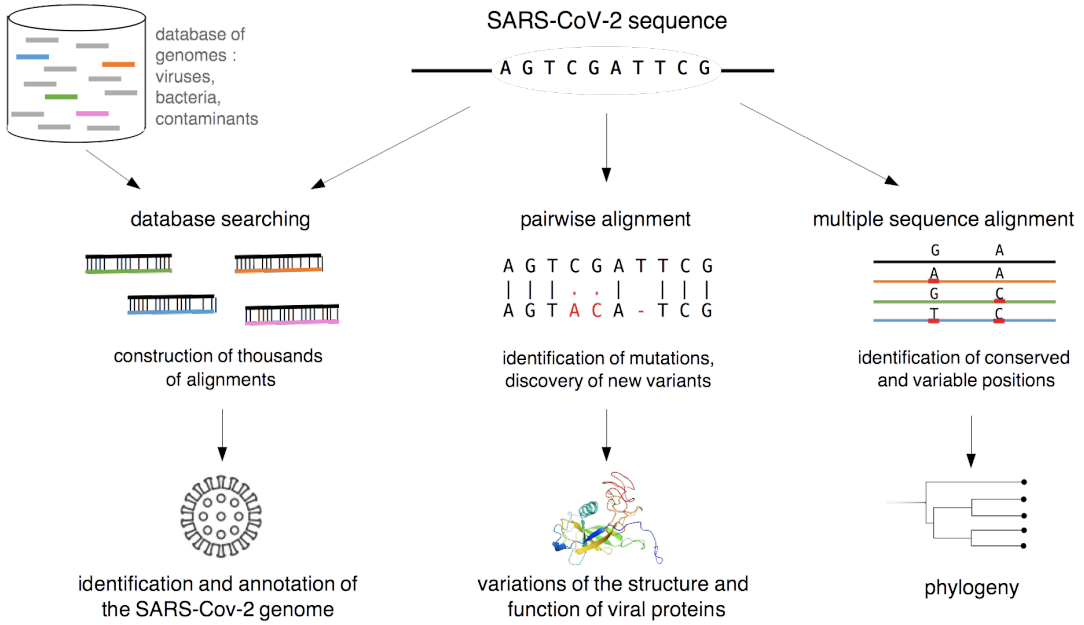
Data filtering. Making sense of the raw reads to obtain the genome of interest requires a series of computational treatments based on string algorithms capable of processing big data efficiently. The first problem is that the sequencing data originates from all microorganisms present in the clinical sample, including possible contamination from the human host. The strategy to overcome this obstacle is to filter out reads from the host background or any other known species. This process is performed by mining large genome databases comprising a full range of known microbes (viruses, bacteria, fungi, and parasites) as well as the human genome. For this, the bioinformatics community has developed genomic search engines, such as Blast [2], that resemble *Google for DNA*. These tools compute *sequence alignments* to distinguish reads that are similar to sequences present in the database from other data (see Box 1). This line of research dates back to the end of the 1990s approximately. The most recent methods have been specifically designed over the last ten years to handle high-throughput sequencing reads. They are able to process gigabytes of DNA sequences in a few minutes. The algorithmic core relies on advanced concepts from information theory, such as compression, hash functions, index data structures [3, 4]. These advances have made it possible to isolate the reads originating from the novel virus from the other reads. This isolated portion typically represents less than 1% of the initial data.

Genome assembly. Once the reads of interest have been isolated, the final step is to assemble them, reconstructing the genome's sequence from the puzzle of reads. De novo assembly of a new genome is like assembling a furniture kit for which the instruction guide has been lost. Additional sources of complexity are that there are hundreds of thousands of small pieces, many of which look alike since they are all written in the same four-letter alphabet, A, C, G, T, and some of which contain erroneous letters due to sequencing errors. De novo assembly is still a major challenge for large genomes, measuring up to billions of nucleotides. But software programs developed in the last decade can now easily solve the assembly puzzle for simple genomes, such as viral ones. State of the art methods rely on De Bruijn graphs [5], for which we give a brief introduction in Box 2. The assembly results in the reconstruction of the virus's genomic sequence: in the case of SARS-CoV-2, it is a single strand RNA sequence composed of approximately 30,000 nucleotides. The first reference sequence was made publicly available on January 12, 2020. All in all, it took less than two weeks to obtain a genome of what is now known as the SARS-CoV-2 virus.

Internet resources

- The first published reference genome is available on the NCBI website with identifier NC_045512: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.

BOX 1. Alignment of biological sequences

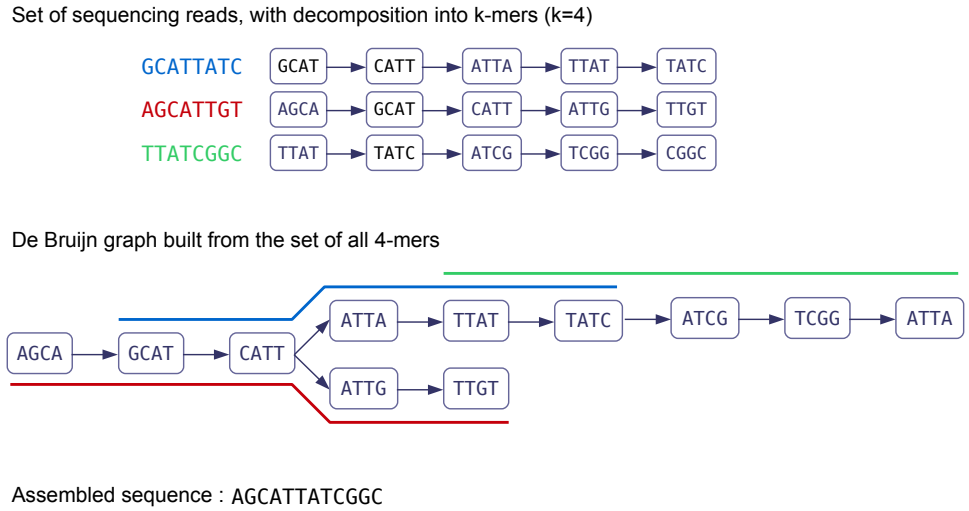


Sequence alignment serves multiple purposes: filtering out the sequencing data, finding genes on the genome, identifying variants between strains, and building multiple sequence alignments, for example.

Alignment is the algorithmic process of comparing sequences to detect similarities and differences [1]. Differences correspond to mutations or sequencing errors: replacement of one nucleotide by another, insertion of an extra nucleotide, or deletion of a nucleotide. When comparing two sequences, the number of all possible alignments is exponential because of the combinatorics of insertions and deletions. Therefore, the naive approach of computing all possible alignments is infeasible in practice. This problem can be best solved as an optimization problem using dynamic programming, a common algorithmic paradigm. It works by dividing the problem into smaller subproblems and is able to compute the optimal alignment without resorting to approximations.

Pairwise alignments can also be computed between a single sequence and a database containing millions of genomes. This is done to remove reads from the human genome (3 billion nucleotides) and other known respiratory parasites when sequencing pulmonary samples. For such large-scale comparisons, computing exact dynamic programming alignments between the query sequence and each of the database sequences would require months of calculations. To overcome this difficulty, bioinformatics researchers have proposed efficient heuristics able to process gigabytes of DNA data with a desktop computer. The search is performed by organizing the database into an index structure that lists all words of a given length, called k-mers, and allows direct access to those words in the genomes. Examples of such indexes are hashtables, or compressed tree-like data structures such as suffix arrays and FM-index. After rapidly identifying the few sequences that share such small similarities with the query sequence, more precise and longer alignments can be performed with the classical dynamic programming paradigm. This kind of approach makes it possible to identify a needle in a haystack: similarities as short as ten nucleotides between the sequence of interest and the database. With big data such as this, short matches can occur by chance with no biological meaning. It is thus crucial to evaluate the statistical significance of the alignments found, which is done with E-value calculations that measure the number of alignments that would be expected due to chance alone. For example, it is possible to find local similarities between the SARS-CoV-2 spike protein gene and organisms in the tree of life as diverse as the two model bacteria *Escherichia coli* and *Bacillus subtilis*, the maize plant, the zebrafish, or even an unrelated virus, such as HIV. Those matches all have E-values greater than 0.01, which is not significant, while alignment between the SARS-CoV-2 spike gene and other coronavirus spike genes reaches an E-value as low as 0.

BOX 2. De Bruijn graphs



The most commonly employed method to reconstruct genomes from sequencing reads relies on *de Bruijn graphs*. The name comes from the mathematician Nicolaas de Bruijn, who introduced these data structures in the 1940s as a combinatorial object. De Bruijn graphs made their entry into bioinformatics 60 years later, with the advent of high throughput sequencing. Indeed, the amount of sequencing data routinely generated by an experiment (gigabytes or even terabytes) made the existing software obsolete and necessitated a new paradigm for genome assembly. When applied to the assembly problem, the principle of De Bruijn graphs is as follows. Reads are split into strings of a particular length k, called *k-mers*, which are shorter than entire reads. In the above example, we have $k = 4$. The graph for the set of reads is then constructed by taking all k-mers as vertices and adding edges between vertices with an overlap length of exactly $k - 1$. The original genome sequence is obtained as a path in this graph.

The main advantage of De Bruijn graphs is that memory usage scales with the number of unique k-mers in the set of reads, rather than the number of reads. Moreover, edges are implicit since they are deduced from the two adjacent vertices. The memory footprint can further be decreased using probabilistic filters, such as the Bloom filters. In real life, the value of k ranges between 20 and 130, depending on the size of the genome. The method is also adapted to handle experimental sequencing data: the existence of erroneous reads due to sequencing errors, the presence of repeated regions in the genome, or the presence of unsequenced regions.

Further reading on sequence alignment and assembly

[1] Sequence Alignment. S.F. Altschul and M. Pop, In: Handbook of Discrete and Combinatorial Mathematics (2017). https://europepmc.org/article/NBK/nbk464187

[2] The BLAST Sequence Analysis Tool. T. Madden, The NCBI Handbook (2013). https://www.ncbi.nlm.nih.gov/books/NBK153387

[3] Genome-Scale Algorithm Design : Biological Sequence Analysis in the Era of High-Throughput Sequencing. V. Makinen, D. Belazzougui, F. Cunial and A.I. Tomescu (2015)

[4] Alignment of Next-Generation Sequencing Reads. K. Reinert, B. Langmead, D. Weese and D.J. Evers, *Annual Review of Genomics and Human Genetics* 16:133-151 (2015) doi.org/10.1146/annurev-genom-090413-025358

[5] How to apply de Bruijn graphs to genome assembly. Ph. Compeau, P. Pevzner & G. Tesler. *Nature Biotechnology* 29 (2011)

Bibliographical sources on sequence alignment and assembly

[6] A new coronavirus associated with human respiratory disease in China. F. Wu et al. *Nature*, 579:265-269 (2020). doi:10.1038/s41586-020-2008-3

[7] Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. LL Ren et al. *Chinese Medical Journal* 133(9):1015-1024 (2020). doi:10.1097/CM9.0000000000000722

[8] Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease 2019. Z. Shen et al. *Clinical infectious diseases*, 71,15, 713-720 (2020). doi:10.1093/cid/ciaa203

[9] RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. L. Chen et al. *Emerging microbes & infections*, 9,1 313-319 (2020). doi:10.1080/22221751.2020.1725399

[10] Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Huang C, Wang Y, Li X, et al. *Lancet* 395:497-506 (2020) doi:10.1016/S0140-6736(20)30183-5

1.2 First hints on the biology of SARS-CoV-2

Knowing the genome sequence is a major milestone in the understanding of a disease. In the first place, it opens the way to phylogenetic analysis of the virus by identifying closely related viruses sharing evolutionary relationships. Indeed, large-scale sequence comparison with viral databases revealed that the newly sequenced genome is a novel betacoronavirus, one of four genera of coronaviruses. It shows more than 85% similarity to several bat-derived coronaviruses, while being more distant from other known human betacoronaviruses. It has 79% similarity with SARS-CoV, responsible for the 2003 outbreak of SARS in Asia, and 50% similarity with the Middle East respiratory syndrome coronavirus, MERS-CoV. The availability of all these related virus genomes makes it possible to formulate preliminary hypotheses on the origin of the virus, which we detail in Section 3. It also allows the characterization of proteins encoded by the genome that govern the functioning of the virus, which we detail below.

A quick reminder on the viral cycle. Viruses cannot function by themselves, and their survival is dependent on host cells. For this, they hijack the host's cell machinery to make copies of themselves and infect other cells. This general life cycle relies on five steps:

- the *attachment*: the virus recognizes and attaches to receptor proteins on the surface of the human cells,
- the *entry* of the virus in the host cell, with the injection of its genetic material,

– the *replication* of the virus in the infected cell: during this stage, the virus synthesizes its proteins with the help of the cell machinery of its host,

– the *assembly* of those proteins to produce new virions,

– the *release* of newly formed virions out of the host cell, causing the cell to burst. The new viral particles are ready to infect other cells to repeat the same cycle.

Each of those steps involves dedicated proteins that take part in the biology of the virus and are potential therapeutic targets. So it is crucial to analyze the genome of the virus in order to identify genes coding for those proteins.

Comparative genomics. One basic approach to gene finding is *homology*. In our case, this principle is applied to compare the newly sequenced genome to other betacoronaviruses for which a genome analysis has already been performed. A significant degree of similarity between two genomes serves as strong evidence to infer that the sequences share a common evolutionary history, and that functional elements found in one sequence should be present in the other sequence with minor changes [11, 12]. This analysis makes it possible to transfer the accumulated knowledge about the genomes of known related viruses to the novel coronavirus. From a computational point of view, genome comparison is performed, once again, using alignment algorithms (see Box 1). In the case of SARS-CoV-2, this search revealed that the new genome contains 27 proteins that are conserved across betacoronaviruses, most of them being found in all coronaviruses. The amino-acid sequences of the proteins are derived from the genomic sequences by *in silico* translation with the genetic code. Among them, one can find four structural proteins that make up the viral particle and are required for the virus to infect cells: the protein *spike* S that mediates virus entry into the host cell – as we shall see in detail in Section 2.1, the small envelope protein E that gives the virion its shape, the nucleocapsid protein N which binds the viral RNA and also interacts with a number of cellular components, and the membrane protein M involved in the assembly and release steps. While being very similar to proteins found in other betacoronaviruses, these structural protein sequences show several local differences that are specific to SARS-CoV-2 and are likely to be one of the causes of the functional and pathogenic divergence of this virus. To measure the potential impact of these differences, one can look at *motifs* present in the protein's amino-acid sequence. Such motifs are regions of the protein that are more specifically involved in the function and the structure of the molecule, and that show a higher degree of conservation between species. Modeling motifs is another way to capture homology. This approach relies on probabilistic models based on Hidden Markov Models, which encapsulate multiple sequence alignments to assign probabilities about the presence of each amino acid in each position of the alignment [13, 16]. Box 3 shows the genome organization of SARS-CoV-2, compared to SARS-CoV and MERS-CoV, and also the Receptor Binding Domain of the protein spike S.

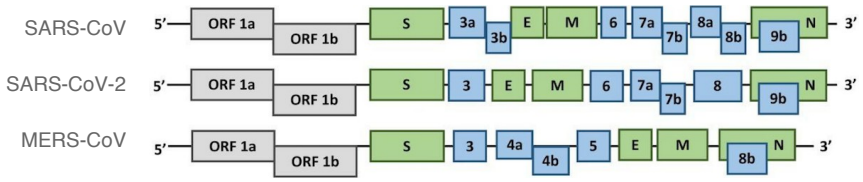
Despite all this information and all these advances, the SARS-CoV-2 genome has not told us all its secrets. For example, homology search with more distant coronaviruses shows the unexpected presence of putative small genes, known as *overlapping genes*, that are hidden behind other genes, meaning that the same portion of the genome can code for distinct proteins [17]. The function of these genes is still an open question.

Internet resources

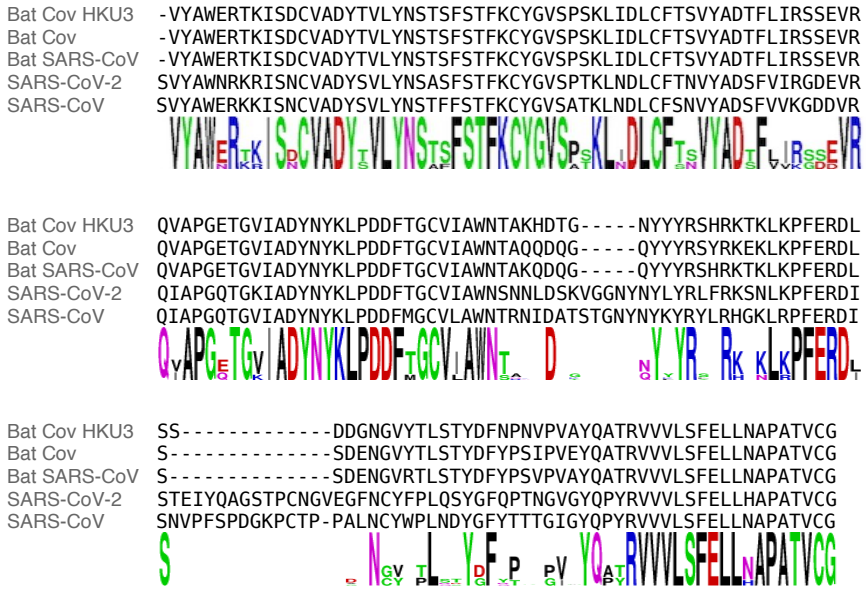
- The UCSC SARS-CoV-2 Genome Browser: `https://genome.ucsc.edu/covid19.html`. This portal is developed by UC Santa Cruz Genomics Institute and provides user-friendly tools to visualize the genome, along with its genes and mutations.
- The Ensembl SARS-CoV-2 assembly and gene annotation resource: `https://covid-19.ensembl.org/Sars_cov_2/Info/Annotation`. On this site, one can find the reference genome, genes, and protein annotations with motifs.

BOX 3. Comparative genomics

(A) Genome organization of SARS-CoV-2



(B) Example of motif: the Receptor Binding Domain of the protein spike S



(A) Genome organization of SARS-CoV-2. ORF1a and ORF1b contain 16 non-structural proteins that are required for replication and transcription. The genes encoding structural proteins spike (S), envelope (E), membrane (M), and nucleocapsid (N) are in green. The genes encoding accessory proteins are in blue. Genomes for two other human betacoronaviruses are also displayed to illustrate the conservation between closely related viruses: SARS-CoV and MERS coronaviruses (figure adapted from [15]). (B) Example of motif: a closer look at the amino-acid sequence of the Receptor Binding Domain (RBD) in protein spike S. This motif, starting at position 349 and ending at position 526 of the protein, attaches to the host receptor, initiating the infection. It is found by comparison with other RBDs in other coronaviruses using a Hidden Markov Model. The top three sequences are from bat betacoronaviruses (Bat Cov HKU3, Bat CoV, and Bat SARS CoV). The two last sequences are SARS-CoV-2 and SARS-CoV. The multiple sequence alignment was built here with Muscle. The RBD is presented in further detail in Section 2.1 and in Box 8.

Further reading on genome and protein analysis

[11] Genome annotation: from sequence to biology. L. Stein, *Nature Reviews Genetics* 2, 493–503 (2001). doi.org/10.1038/35080529

[12] Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. G.F. Ejigu and J. Jung, *Biology* 9, 295 (2020) doi.org/10.3390/biology9090295

[13] What is a hidden Markov model? S. Eddy, *Nature Biotechnology* 22, 1315–1316 (2004). doi.org/10.1038/nbt1004-1315

Bibliographical sources on genome and protein analysis

[14] Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. A. Wu et al., *Cell Host Microbe* 11;27(3):325-328 (2020). doi:10.1016/j.chom.2020.02.001

[15] A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses. SY Fung et al. *Emerging Microbes & Infections*, 14;9(1):558-570 (2020). doi:10.1080/22221751.2020.1736644

[16] Pfam: The protein families database in 2021. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman, *Nucleic Acids Research* (2020) doi:10.1093/nar/gkaa913

[17] Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. C.W. Nelson et al. *Elife* 1;9:e59633 (2020). doi:10.7554/eLife.59633

1.3 Genomics for public health strategies

In addition to its significance in biological studies, the identification of the reference genome had a direct impact on public health policies, namely with two essential tools to track the spread of the disease: (1) the development of massive and low-cost diagnostic testing facilities with *PCR tests*, and (2) the genomic surveillance of the virus to monitor its dynamics and its mutations with *large-scale sequencing of individuals*.

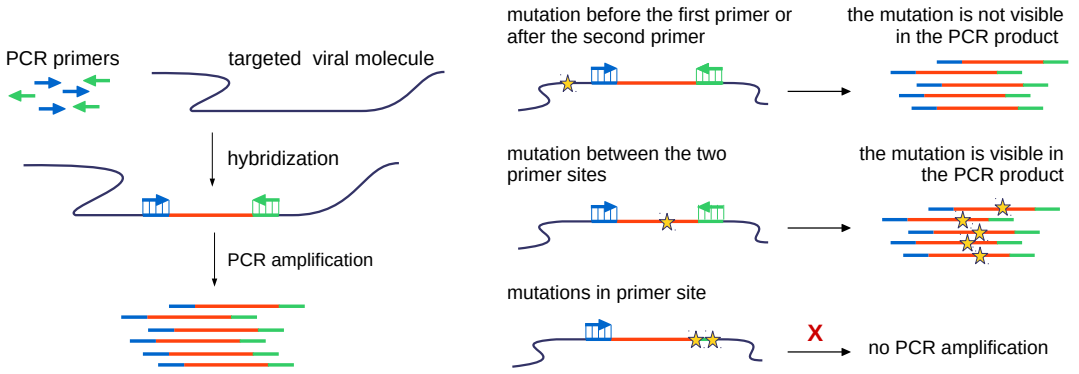
Widespread laboratory testing. At the onset of the outbreak, there was a need to identify people carrying the virus, either symptomatic or not, in order to treat and isolate them to stop the spread of the virus. The first tests were designed from the genome sequence using a well-known low-cost and robust molecular methodology: the PCR (Polymerase Chain Reaction). PCR can indicate whether the virus's genome is present in a population of cells without sequencing the entirety of the genetic material. The tested cells come from respiratory secretions, such as those collected by nasopharyngeal swab samples. PCR tests had been previously developed for other coronaviruses, such as SARS-CoV and MERS-CoV, among other infectious agents, so it was natural that it became the first and preferred testing method for SARS-CoV-2 diagnosis.

The principle of PCR design is to identify small portions of the virus genome, called *primers*, that can characterize this genome. In other words, it amounts to finding specific patterns that delineate a representative region of the genome of interest. Bioinformatics guidelines and computer programs for designing effective PCR primers have been available since the early 2000s, as explained in Box 4. The first PCR tests for SARS-CoV-2 were designed in February 2020 from 95 genomes, mainly for research purposes [18]. Commercial kits were available for widespread laboratory testing as of the beginning of Summer 2020. In Europe alone, millions of tests are now performed each week.

BOX 4. Bioinformatics for PCR design

(A) Principle of PCR amplification

(B) Impact of mutations



PCR (polymerase chain reaction) is a widely used molecular technique that allows one to rapidly generate a high number of copies of a specific DNA or RNA strand measuring up to a few hundred nucleotides long. In the case of an RNA sequence, such as the SARS-CoV-2 genome, the correct name is RT-qPCR, where RT stands for reverse transcriptase, the process of moving from RNA to DNA, and q for quantitative. PCR has multiple applications. For SARS-CoV-2, PCR is used both for test design and tiled amplicon sequencing, as described in Subsection 1.3.

The general principle of PCR consists of repeatedly amplifying the genetic material so that the amount of DNA can be quantified. It does not amplify the whole sequence, but only a region of interest. For this purpose, a PCR is based on two short DNA sequences, called *primers*, delimiting the region to be amplified. These primers, about 20 nucleotides long, are complementary to the target genome and need to be specific to the sequence of interest such that only it is amplified. Thus, the first challenge is to conceive suitable PCR primers for targeting the desired region of the SARS-CoV-2 genome without inadvertently targeting other regions, or, for that matter, other pathogens or human genes present in a given nasal sample. This is achieved by aligning the potential primers on known pathogens and human genomes to be sure that there is neither cross-hybridization nor similarity with any other organism. Another challenge is designing robust primers that could attach solidly to the genome strand. Affinity depends on the primer's sequence composition, structure, and dynamics (see Box 7). After all these *in silico* steps, the primers should be assessed on real samples showing the effectiveness of the amplification tests *in vitro*.

Once the primers have been validated, their validity must be assessed against new variants of the virus as these appear. Indeed, as the virus has spread around the world, several different lineages have proliferated, each with its specific mutations. There is the risk that existing primers may not be adapted to a new strain if mutations modify the site recognized by a primer, which may prevent a proper amplification. Until July 2020, sequence alignments performed on thousands of SARS-CoV-2 genomes showed that only a few minor strains (about 1%) had pointwise mutations at the location of the primers [19]. However, this is no longer true with the British variant B.1.1.7, which exhibits one mutation on a primer site. In general, it is recommended that the pair of primers can resist at least one punctual mutation. This should be taken into account to prevent loss of hybridization with the target genome or undesired cross-hybridization. Another recommendation for designing commercial PCR tests is to target several regions, and thus design several pairs of primers, in order to enhance robustness. The amplified regions also differ according to the test manufacturer, meaning that some tests can remain effective, while others become obsolete.

Beyond the RT-qPCR, which only quantifies some genetic material, we can access the content of the amplified nucleic sequences with high-throughput sequencing. As this process is massively automated, we can pool tens of thousands of samples and sequence them all together, making testing much faster. In this approach, A DNA tag is added to each molecule, allowing individual samples to be distinguished. Next, efficient bioinformatics methods (partially relying on alignments) are used to identify the sample of origin, and to identify the sequences that have been sequenced. As far as we know, such methods have not been used at a large scale yet.

Towards global genomic surveillance. The availability of the SARS-CoV-2 reference genome is also a powerful tool to establish worldwide surveillance based on large-scale sequencing, such as what has been done in the last few years with Influenza or the seasonal flu, Ebola, and Zika. Indeed, knowing the genome allows for simplified sequencing protocols with substantially reduced costs at an accelerated pace. Such protocols make it possible to collect and analyze numerous SARS-CoV-2 genomes—from different global locations and at different time points—in order to monitor the disease as closely as possible.

We have seen in Subsection 1.1 that the acquisition of the first Wuhan genomes required sequencing the whole RNA content of patients' pulmonary samples, while the viral fraction contained in such samples is extremely small. This produced large amounts of sequence data which necessitated complex downstream bioinformatics analyses. Once the viral genome is known, sequencing efficiency is improved by exclusively targeting the sequences of interest. Tiled amplicons implement such a strategy [20]. Like previously described diagnostic tests, amplicon sequencing relies on PCR amplifications based on the design of a series of primers present in the target genome (as detailed in Box 4). The goal here is to tile the coronavirus genome by short regions surrounded by pairs of primers, each of them being amplified and sequenced as a PCR product. This is a nice combinatorial problem: find a set of primer pairs in the reference genome such that they adequately cover the whole genome and are suitable for PCR amplification. For example, it is possible to tile the genome of SARS-CoV-2 with 137 PCR segments, each 400 nucleotides long, or 299 PCR fragments, each 200 nucleotides long [20]. Once all amplicons are generated, the genome assembly of the newly sequenced viral strain is also made easier by relying on the already known SARS-CoV-2 genome sequence, which serves as a template. De novo assembly is no longer required, and reads are simply aligned to the reference genome (see Box 1). A consensus sequence can then be computed by selecting, at each position, the character observed in the majority of reads aligned at the given position. Sequence variations, such as mutations, are then identified between the newly sequenced viral strain and the Wuhan genome or between different strains co-existing in the sample.

In December 2020, less than one year after the publication of the first Wuhan genome, more than 300,000 other SARS-CoV-2 genomes have been sequenced and assembled by various laboratories and institutions from countries all around the world. This extensive sequencing effort allows us to understand how the virus is evolving, to track the mutations in real time, and to identify emergent strains. One remarkable aspect of this research is that several national and international initiatives have rapidly developed dedicated web portals in order to store this important information and make it freely available on the Internet. This is the spirit of open science. Generalist sequence repositories such as Genbank hosted by the NCBI (American) or the European Nucleotide Archive hosted by the EBI (European), which have been organizing public-domain sequence data sharing for several decades, have developed specific databases and tools for the SARS-CoV-2 data. The GISAID Consortium also provides an essential resource for SARS-CoV-2 genomes. The GISAID database had 339 genomes available at the end of January 2020, and this number grew rapidly, reaching around 80,000 in August 2020 and more than 600,000 in February 2021. Scientists can freely submit sequence data to such repositories, but all data are carefully quality-checked and re-annotated before being publicly shared. The available genomic sequences also come with additional information, such as the sample origin and time, molecular sequencing protocols, patient clinical information, etc. These metadata are well-structured in databases for efficient queries and downstream comparative analyses from this huge collection of sequences.

All this information—genomes and metadata—is used to study the propagation of the virus around the world. It is essential to identify mutations that could impact its pathogenicity and its transmissibility. In this perspective, the UK was able to detect the B.1.1.7 lineage early, in October 2020, and hence monitor the emergence of what is now commonly referred to as the *British variant*. This was only made possible thanks to an ambitious sequencing policy. Indeed, as of December 2020, more than 120,000 genomes have been published in the UK, compared to less than 3,000 in France in the same period. Another variant, B.1.351, referred to as the *South African* variant, was also detected in October 2020 and quickly became the dominant strain in that country. Sequencing shows that in this variant, several mutations are located in the gene coding for the spike protein S, a key protein involved in the virus's entry into human cells, which is consequently targeted by vaccines (as we will see in Section 2.1). This

observation indicates that B.1.351 should be closely examined to check whether it could escape vaccine-induced protection. This case provides critical evidence of the public health benefits of mass sequencing. Finally, in a broader perspective, the availability of a large number of genomes is a prerequisite for modeling the evolution of the virus. We will explain this aspect in more detail in Section 3.

Internet resources

- European Centre for Disease Prevention and Control, and more specifically data on testing for COVID-19 by week and country: https://www.ecdc.europa.eu/en/publications-data/covid-19-testing
- GISAID, https://www.gisaid.org, was created in 2008 to ensure rapid sharing of data from influenza epidemics and is now a key resource for COVID-19 genomes. It includes genetic sequences and related clinical and epidemiological data associated with human viruses, and geographical as well as species-specific data associated with avian and other animal viruses, to help researchers understand how viruses evolve and spread during epidemics and pandemics.
- NCBI COVID portal: https://www.ncbi.nlm.nih.gov/sars-cov-2
- EBI COVID portal: https://www.covid19dataportal.org

Bibliographical sources on genome and protein analysis

[18] Primer design for quantitative real-time PCR for the emerging Coronavirus SARS-CoV-2. D. Li, J. Zhang and J. Li, *Theranostics* 10(16):7150-7162 (2020). doi:10.7150/thno.47649

[19] Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT–qPCR primer–probe sets. C.B.F. Vogels, A.F. Brito, A.L. Wyllie et al. *Nature Microbiology* 5:1299–1305 (2020). doi.org/10.1038/s41564-020-0761-6

[20] Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. M. Xiao et al., , M., Liu, *Genome Medicine* 12, 57 (2020). doi.org/10.1186/s13073-020-00751-4

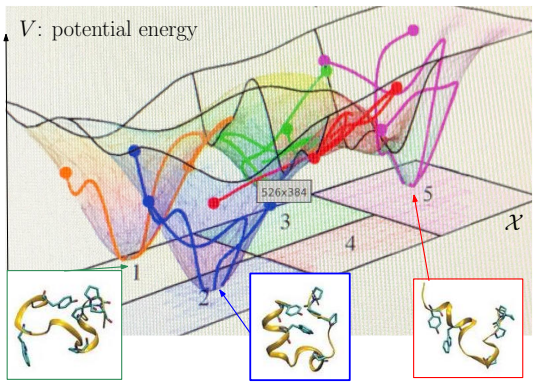
2 Fighting the disease, advances in health care

The establishment of the SARS-CoV-2 genome alone is not sufficient to understand its pathogenicity and then develop vaccines and treatments. The management of the disease also requires modeling the mechanisms of transmission and infection of the human body.

The virus's biological functions rely on the formation of molecular complexes, such as protein complexes, which are themselves generally involved in larger interaction networks and pathways. These complexes and pathways are closely related to the several steps of the viral cycle mentioned in Section 1.2: first, the virus's entry into the human cell, then the production of viral proteins and formation of new virions, and finally, the release of virions out of the host cell. Each of these steps is naturally critical for the success of the infection, and understanding the corresponding mechanisms at both the molecular and systems level is key to combating the virus.

For example, most of the COVID-19 vaccines, such as the Pfizer-BioNTech or Moderna vaccines, focus on the viral gene that codes for the spike protein and plays a key role in the entry step of the virus. In this context, the protective immunity induced by the vaccine comes from the fact that the host antibodies that 'learn' to recognize the spike protein should also be able to neutralize the virus. So, a nuanced understanding of the structural interactions involving the spike protein is needed to explain the particular problem of the virus entry. In Section 2.1, we explain how *structural bioinformatics* provides computational methods for this task.

BOX 5. Potential Energy Landscape of a biomolecular system



The energy landscape of a biomolecular system encodes its structural, thermodynamic, and kinetic properties. The potential energy landscape (PEL) associates a potential energy to each conformation. The main challenges of molecular simulation are to (1) find significant local minima of the PEL, (2) compute statistical weights of catchment basins by integrating Boltzmann's factor (for a so-called NVT thermodynamics ensemble), and (3) identify transitions. Practically, the conformational space $\mathcal{X}$ has dimension $d = 3n$, with n the number of atoms.

Another illustrative case is the *cytokine storms* occurring in some patients, particularly in the lungs. This phenomenon is a cascade of exaggerated responses of the host's immune system, characterized by sudden and massive releases of cytokines. Although these proteins are a normal part of the body's immune response to infections, their excessive production can cause life-threatening symptoms. Understanding and modeling the dynamics of such complex interactions raises difficult *systems biology* issues. We elaborate on these issues in Section 2.2, focusing on universal questions and a range of biological systems.

2.1 On the role of structural models in combating the virus

Structural biology aims to bridge the gap between the spatial conformation and dynamics of biomolecules (mainly proteins and nucleic acids) and their function. To understand the associated computational challenges, observe that a molecule with n atoms is described by $3n$ Cartesian coordinates defining its *conformational space*. As with any physical system, a biomolecule and its environment (i.e. the solvent) can be associated with a *potential energy*. When chemical bonds are not altered, this potential energy is computed from molecular mechanics-based force fields, which rely on classical mechanics. The graph of the potential energy defines the *potential energy landscape* (PEL), described in Box 5 [21]. The sheer difficulty of developing accurate models using PEL comes from two sources. First, biomolecular systems are inherently large (tens of thousands of atoms), yielding very high dimensional PEL. Second, molecular motions span ~ 15 and ~ 4 orders of magnitude in time and amplitude, respectively [22]. For example, the vibrations of atoms sharing a covalent bond occur on the femtosecond time-scale, while biologically relevant time scales are beyond milliseconds. To study such phenomena, a mix of experiments and modeling-simulation techniques are used.

From the experimental standpoint, the structures of biomolecules are obtained thanks to a variety of experiments, notably X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy (cryo-EM), which deliver atomic models (i.e. atomic coordinates) of key configurations. In the case of SARS-CoV-2, as of December 2020, circa 650 structures have been deposited in the Protein Data Bank. However, the main mechanisms of interest are highly dynamic, and the structures obtained are *merely* snapshots in a complex *movie*.

From the modeling standpoint, the *structure* of a macromolecular system (isolated molecule or complex) requires the characterization of active conformations and important intermediates in functional pathways. In assigning occupation probabilities to these conformations, one deals with *thermodynamics*. Finally, transitions between the states, modeled by a master equation (a continuous-time Markov process), correspond to *kinetics*. These concepts have a direct translation on the PEL: stable states correspond to *significant* basins of the PEL; thermodynamics require integrating Boltzmann's distribution on the basins; and finally, kinetic models qualify the dynamics between basins. As should be clear from the description of PEL above, these questions require exploring and characterizing very high dimensional spaces over time scales which span circa 15 orders of magnitude. The mathematical and computational questions faced are immense.

The problems discussed above are essentially open, as experiments and simulations are currently unable to unveil structural, thermodynamic, and kinetic properties of large or dynamic systems. Fortunately, combining experimental and modeling approaches can provide invaluable insights. This has been the case with SARS-CoV-2. Before presenting selected technicalities, let us briefly inspect the mechanisms of infection by the virus, and more specifically, the entry of the virus into the host cell, which makes it possible for the virus to inject its genetic material into the cell and replicate itself.

A closer look at the entry step of SARS-CoV-2. As in all molecular mechanisms, the entry step involves a mutual recognition between two proteins, one from the virus and one from the infected cell, followed by a highly dynamic event.

On the virus side, we have already seen that SARS-CoV-2 infects human cells with *spikes* found on its envelope. These spikes are homotrimers (3 copies) of the protein S [27, 28]. Each chain consists of two domains, S1 and S2, separated by cleavage sites denoted S1/2 and S2' [29, 30, 31]. Domain S1 contains the *Receptor Binding Domain* (RBD), while domain S2 contains the *fusion machinery*. Membrane fusion is an essential process involved in trafficking between cells and cellular compartments, the exchange of genetic information across individuals, and also in viral infection. Fusion is accomplished by *fusion proteins*, which were first discovered in enveloped viruses. There are three classes of such proteins, namely class I (found, for example, in influenza and SARS-CoV-2), class II (e.g. in dengue), and class III (e.g. in herpes viruses). The main steps of the mechanism are shown in Box 6: (1) attachment of the RBD to its target on the cell to be infected, (2) proteolysis cleavage/activation at S1/S2, triggering the release of the S1 subunit, (3) second cleavage at S2', triggering fusion machinery refolding—the anchoring of the fusion peptide into the target membrane—and the envelope-membrane fusion. The virus genome delivery into the target cell follows.

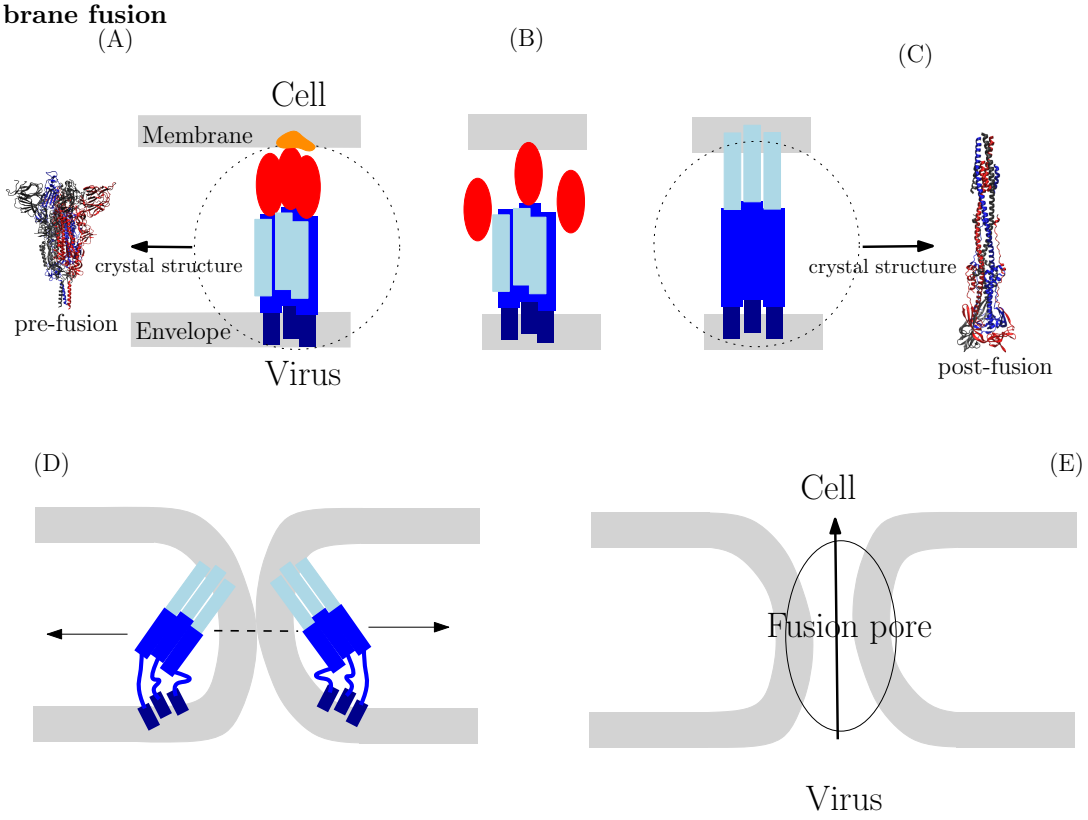
On the host side, in human cells, SARS-CoV-2, like other betacoronaviruses such as SARS-CoV, targets the protein ACE2 [32], a membrane-bound enzyme catalyzing the hydrolysis of angiotensin II into angiotensin.

Interestingly, the RBD adopts two conformations which are termed up/down or closed/open. The RBD of SARS-CoV-2 stands up less often than that of SARS-CoV, which likely favors evasion from the immune system [33]. This behavior also disfavors binding to the target (binding in the down configuration yields steric clashes), but the lesser proportion of RBD standing up is counterbalanced by a high affinity for ACE2 [33].

The mechanism sketched out above offers two main *therapeutic opportunities*. The first one consists of preventing the attachment of the RBD to its receptor, which may be done by eliciting an immune response that provides antibodies to block the RBD, or by providing drugs achieving the same effect. The second one relates to the ability to block the refolding of the fusion machinery prior to membrane fusion. We detail each route below.

Designing competitive blockers targeting the spike of SARS-CoV-2. The first strategy to prevent infection by SARS-CoV-2 is to block its RBD. This can be done by mimicking the region of ACE2 found at the interface with the RBD, as shown in Box 8(A). This goal was recently pursued using *de novo* designed mini-proteins of circa 50 amino acids [34]. From the experimental standpoint, a candidate protein can be validated in two complementary ways: first, by measuring its binding affinity

BOX 6. SARS-CoV-2: host cell entry mechanism via virus envelope - cell membrane fusion



The complete spike involves two domains, S1 and S2. S1 contains the receptor-binding domain (RBD, red ellipsis), while S2 contains the fusion machinery (blue rectangles). **(A)** Attachment of the RBD to its receptor ACE2 (orange molecule, shown only once to avoid clutter) **(B)** Cleavage step removing the S1 subunit **(C)** Refolding of the fusion machinery, and anchoring into the cell membrane **(D)** The endpoints of the fusion protein bring the viral envelope and the cell membrane into close proximity **(E)** The *collaboration* between several fusion proteins triggers the formation of the hemi-pore and pore via virus envelope - cell membrane fusion. The virus can inject its genetic material. Figure adapted from [25]. Structures displayed on the left and right-hand sides: PDB:6xr8 and 6xra, from [39].

BOX 7. Binding affinity

Consider two proteins P and L which associate to form the complex PL, which itself dissociates into P and L:

$$P + L \rightleftharpoons PL. \tag{1}$$

These two processes (association and dissociation), illustrated by the double harpoon, are due to opposing phenomena: on the one hand, attraction forces result in a decrease of the potential energy when P and L bind; on the other hand, thermal fluctuations result in dissociation of the complex PL. The chemical equilibrium [23] is qualified by the association constant $K_a = [PL]_{Eq}/[P]_{Eq}[L]_{Eq}$ computed from the equilibra concentrations of the three molecular species. Equivalently, one can use the dissociation constant $K_d = 1/K_a$. The dissociation constant is also related to the variation of the *Gibbs* free energy by

$$\Delta G_a^0 = RT \log {}^{K_d}/_{c^0}. \tag{2}$$

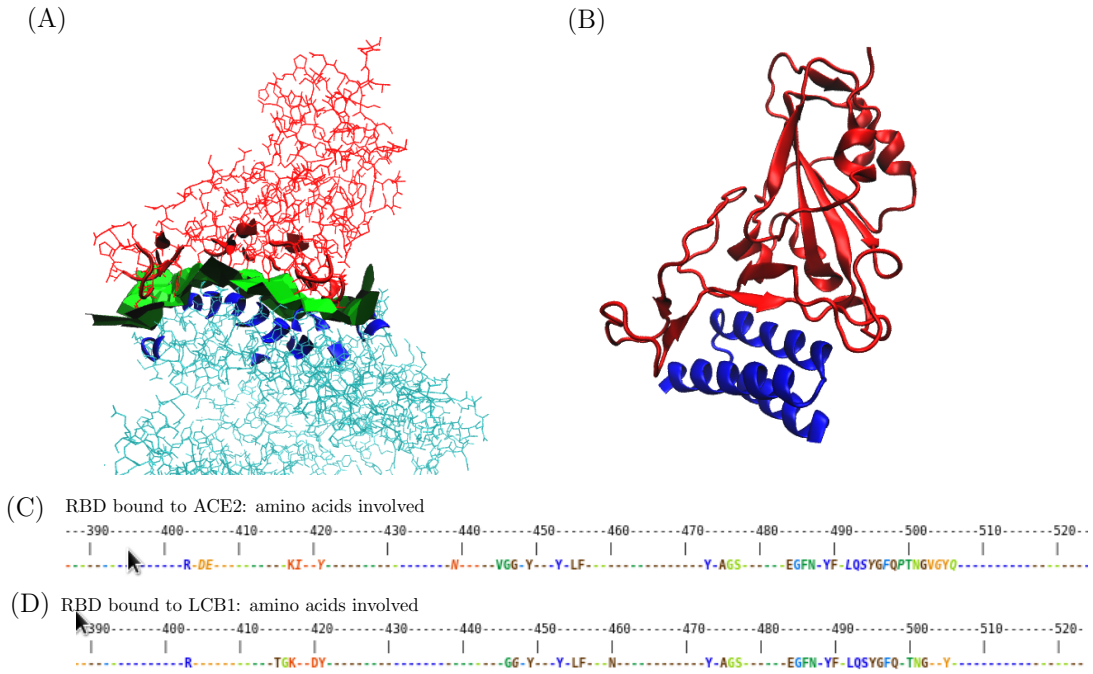
For biological complexes, K_d spans a wide range of scales: from 10^{-6} (signaling protein), to 10^{-12} (small molecules inhibiting proteins), and even 10^{-15} (biotin-avidin complex, strongest known non-covalent interaction). Estimating ΔG_a^0 or equivalently K_d is key to qualify the stability of a complex. In theory, the calculation can be carried out using the potential energy of the system [24]. In practice, such calculations are currently intractable, so that a whole hierarchy of approximations are resorted to.

(K_d) with the target (here the RBD), and second, by solving the structure of the complex via X-ray crystallography or cryo-EM. However, 50 amino acids yield the astronomical number of $21^{50} \sim 10^{66}$ candidate molecules. A careful in silico design is therefore mandatory, based on three steps: first, a selection of suitable amino-acids based on those found in the protein ACE2; second, several simulation rounds to check whether the two molecules bind, using *docking* and *scoring* methods [35]; and third, a careful assessment of the interface in the docked complexes. The reader is referred to [34] for the protocol used for the first two steps. In the sequel, we illustrate a recently developed tool to analyze the interface between the RBD and a candidate blocker–the reader is referred to [36] for full details.

Given a protein-protein complex, a Voronoi model can be used to find the amino acids of the two partners defining the *interface* (Box 8(B)). These amino acids can then be pulled back onto the sequence, defining an *interface string*. Moreover, performing a multiple sequence alignment of several such strings yields a *multiple interface string alignment*, or MISA. This construction can be done for two complexes, namely (i) the RBD and ACE2, and (ii) the RBD and a candidate blocker. For the example used in our illustration (LCB1, [34]), remarkably, it is seen that both the blocker and ACE2 target the same amino acid on the receptor binding domain (RBD) (Box 8(C,D)). Beyond this particular case, efficient screening identified seven designs with dissociation constant K_d in the range 1-20 nano-molar were obtained, and two with $K_d < 10^{-9}$ [34] (Box 7 on binding affinity). These molecules also blocked SARS-CoV-2 infection in culture cell lines, and clearly provide a sound starting point to design anti-SARS-CoV-2 therapeutics.

Delineating the fusion machinery refolding of SARS-CoV-2. The second strategy to prevent infection by SARS-CoV-2 is to preclude the refolding of its fusion machinery, which hinders the fusion between the viral envelope and the cell membrane. As noted above, this process only concerns the S2 domain of the spike –the cleavage steps remove S1. Interestingly, a similar mechanism holds for the influenza virus, whose class I fusion, called hemagglutinin, contains two domains, HA1 and HA2, equivalent to the S1 and S2 domains in SARS-CoV-2. The example of HA2, which has been under scrutiny for decades, is very informative. In particular, it is known that so-called broadly neutralizing antibodies bind to the stem of HA2, preventing its refolding and therefore fusion [37]. Interestingly, the refolding of the region aa 33-172 of HA2 (140 a.a., 1150 heavy atoms) was recently studied [38] in the framework of PEL. A systematic exploration and characterization of the PEL of this system resulted in a database with $\sim 33,000$ local minima and $\sim 41,000$ transition states. These numbers show the complexity of such dynamical processes. Importantly, the most stable structures (identified by their

BOX 8. Designing miniproteins inhibiting the entry of SARS-CoV-2 by blocking its Receptor Binding Domain (RBD)



(A) (PDB: 2ajf) Complex between the RBD of SARS-CoV-2 (red) and ACE2 (cyan). The Voronoi interface model in green identifies the amino acid at the interface of the complex. To each Voronoi tile in green, corresponds a pair of atoms, one on each partner, in direct contact in the complex. The corresponding amino acids are displayed in cartoon mode (solid blue, red). On ACE2, one clearly distinguishes contributions from one long helix. The set of amino acids found at this interface, once pulled back onto the protein sequence, defines an *interface string* summarizing the interface. Aligning such strings defines a *multiple interface string alignment* or MISA. **(B)** (PDB: 7jzu) Complex between the RBD of SARS-CoV-2 (red) and the designed protein (LCB1) meant to compete with ACE2. Note the similar helical structure. **(C)** MISA of the RBD bound to ACE2 (complex in panel (A)) **(D)** MISA of the RBD bound to a candidate blocker (complex in panel (B)) Note that the *footprints* of ACE2 and LCB1 on the RBD are highly similar. Panels (C,D) prepared using the method from [36].

statistical weights) are candidate conformations that may be targeted by therapeutics.

In coronaviruses, structural rearrangements of domain S2 have been characterized experimentally [29, 39], showing that the postfusion S trimer adopts a 180 Ålong cone-like shape. Intuitively, the S2 domain behaves like a *spring-loaded* mechanism, doubling its length along the refolding process. This is an astonishingly large conformational change. While the first and last frames of this *movie* have been solved experimentally, the intermediate meta-stable states are unknown.

Delineating this mechanism, in a manner analogous to the work carried out for influenza HA2, will pave the way to design blockers.

Internet resources

- The Protein Data Bank: `https://www.rcsb.org` and more specifically the COVID-19/SARS-CoV-2 page: `https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true`

Further reading on structural bioinformatics

[21] *Energy Landscapes*. D. J. Wales, Cambridge University Press (2003)

[22] Molecular dynamics: survey of methods for simulating the activity of proteins. S.A. Adcock and A.J. McCammon, *Chemical reviews*, 106(5):1589–1615 (2006)

[23] *Statistical mechanics: entropy, order parameters, and complexity* J. Sethna et al. volume 14, Oxford University Press (2006)

[24] Calculation of protein-ligand binding affinities. M.K. Gilson and H-X. Zhou, *Annual review of biophysics and biomolecular structure*, 36(1):21 (2007)

[25] Viral membrane fusion. S.C. Harrison, *Virology*, 479–480:498–507 (2015)

[26] New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. M. Cao, C. M. Pietras, X. Feng, K. J. Doroschak, T. Schaffner, J. Park, H. Zhang, L. J. Cowen, and B. J. Hescott. *Bioinformatics*, 30(12):i219–i227 (2014)

Bibliographical sources on structural bioinformatics

[27] Assembly of coronavirus spike protein into trimers and its role in epitope expression. B. Delmas and H. Laude, *Journal of virology*, 64(11):5367–5375 (1990)

[28] Structure, function, and evolution of coronavirus spike proteins. F. Li, *Annual review of virology*, 3:237–261 (2016)

[29] Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. A. C. Walls, M. A Tortorici, J. Snijder, X. Xiong, B.-J. Bosch, F. A. Rey, and D. Veesler, *Proceedings of the National Academy of Sciences*, 114(42):11157–11162 (2017)

[30] Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, *Cell* (2020)

[31] Structural basis for potent neutralization of betacoronaviruses by single-domain camelid antibodies. D. Wrapp, D. De Vlieger, K. Corbett, G. Torres, N. Wang ad W. Van Breedam, K. Loes van Schie, M Hoffmann, S. Pohlmann, B. Graham, N. Callewaert, B. Schepens, X. Slelens, and J. McLellan, *Cell* (2020)

[32] Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, *Science*, 367(6485):1444–1448 (2020)

[33] Cell entry mechanisms of SARS-CoV-2. J. Shang, Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach and F. Li, *PNAS*, NA(NA):1–8, (2020)

[34] L. Cao, I. Goreshnik, B. Coventry, J.B. Case, L. Miller, L. Kozodoy, R. Chen, L. Carter, A. Walls, Y-J. Park, E-M Strauch, L. Stewart, M.S. Diamond, D. Veesler, and D. Baker, De novo design of picomolar sars-cov-2 miniprotein inhibitors. *Science*, 370(6515):426–431, (2020)

[35] Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: Capri 7th edition. M. Lensink, N. Nadzirin, S. Velankar, and S. Wodak. *Proteins: Structure, Function, and Bioinformatics* (2020)

[36] Boosting the analysis of protein interfaces with multiple interface string alignment: illustration on the spikes of coronaviruses. S. Bereux, B. Delmas and F. Cazals, *Submitted* (2020)

[37] Broadly neutralizing antiviral antibodies. D. Corti and A. Lanzavecchia, *Annual review of immunology*, 31:705–742 (2013)

[38] Energy landscape for the membrane fusion pathway in influenza a hemagglutinin from discrete path sampling. D. J. Wales, D. F. Burke, and R. G. Mantell, *Frontiers in Chemistry*, 8:869 (2020)

[39] Distinct conformational states of SARS-CoV-2 spike protein. Y. Cai, J. Zhang, T. Xiao, H. Peng, S. M. Sterling, R. M. Walsh, S. Rawson, S. Rits-Volloch, and B. Chen, *Science*, 25(369):1586–1592 (2020)

2.2 Systems-level graphical and executable models

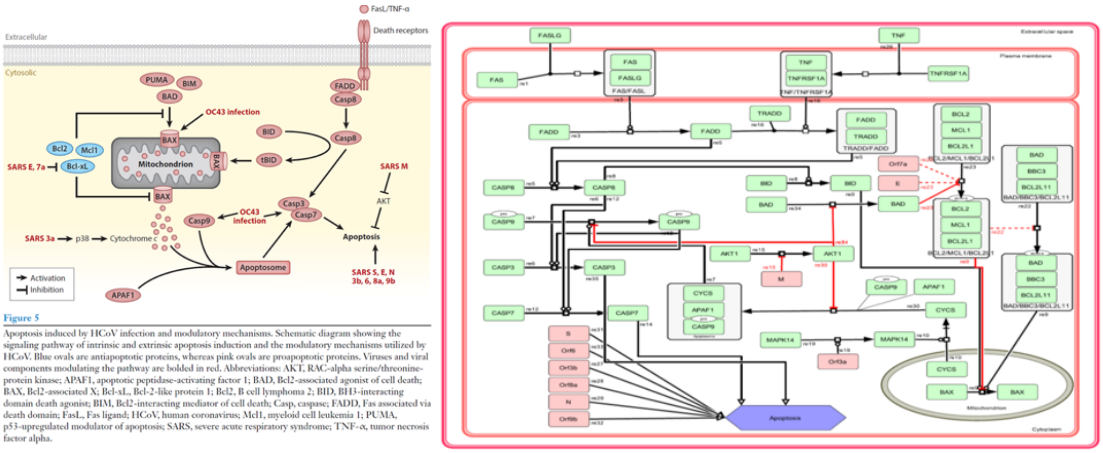
Computational systems biology aims to study interaction mechanisms between various biological entities, such as proteins, genes, RNAs, metabolites, and small molecules like ATP and calcium ions, using computational models. The various biological entities communicate and interact with each other, forming intertwined networks that give rise to the behavior of the system. In a complex biological system like a cell, the emergent behavior is much more complicated than the sum of its subparts. Understanding and elucidating the complex behaviors of biological systems requires appropriate modeling tools and methods [40]. The development of the field of *systems biology* was initiated in the late 1990s in response to the rapid accumulation of biological data, including genome sequences, gene expression data, and protein-protein interactions, for example [41]. The amount of data and the rhythm of its production has only increased since, and integrative, holistic methodologies that can explain the data from a systems perspective are needed to put the bits and pieces of the new knowledge together. The goal is to construct computational models that can mimic the biological system's behavior as realistically as possible, allowing for *in silico* simulations and experiments. Computational systems biology can offer a system-level view, combining structural and dynamical analysis with the integration of multiple layers of biological information. This approach accelerates scientific research and discovery as with the aid of computers, hundreds of experimental conditions can be simulated in a relatively short amount of time. The same scenarios could otherwise take years to be performed in classical experimental lab settings.

The models can be static representations of biological mechanisms (graph-based models), focusing on topology and graph properties, such as connectivity or centrality. They can also be of a dynamic nature, using mathematical descriptions of the regulations. Quantitative, kinetic models and qualitative, logic-based models are two of the main types currently used to model biomolecular networks. These models are used to simulate different conditions, like the presence of a viral protein inside a host cell, or the impact of a drug in a patient [42].

The SARS-CoV-2 pandemic prompted a crisis response, bringing scientists together in new collaborative contexts. In bioinformatics, a range of specialists have contributed their complementary skills to rapidly develop computational pipelines and build graphical and executable, dynamic models spanning many biological processes. In the next part, we will present such efforts to build novel methodologies or adapt existing ones to tackle a variety of questions such as the dynamics of infection, differences between the dynamics of SARS-CoV-2 and other coronaviruses, drug repurposing and identification, and characterization of SARS-CoV-2 virus-host interaction mechanisms.

Building molecular maps of virus-host interaction mechanisms. The COVID-19 Disease Map project [46] is a large-scale, international and interdisciplinary community effort (Luxembourg, France,

BOX 9. Example of an SBGN diagram



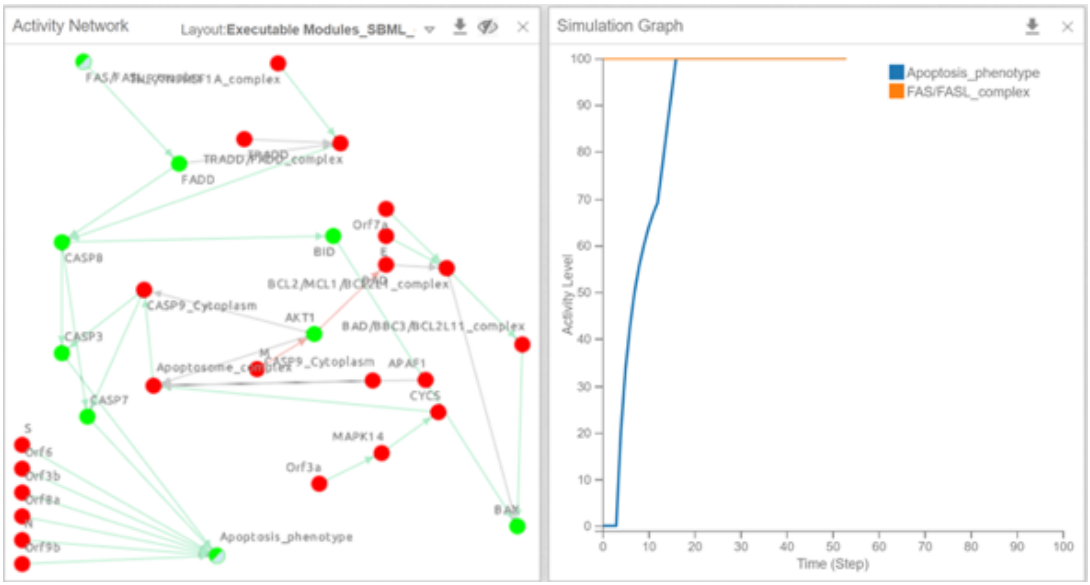
A textbook diagram (left) and an SBGN diagram of the same pathway (right). These represent the Apoptosis diagram featured in Fung and Liu's seminal review, Human Coronavirus: Host-Pathogen Interaction, published in 2019 in the Annual Review of Microbiology [54]. The textbook diagram describes the Apoptosis pathway, along with potential intervention points of the viral proteins with the host cell machinery. As it was published in 2019, the diagram uses information mainly derived from SARS-CoV literature. In the right-hand diagram, the Apoptosis map built with CellDesigner graph editing software [55] is structured into three compartments, namely, the extracellular space containing the ligands, the plasma membrane with receptor-ligand complexes, and the cytoplasm with all signaling and viral proteins. Green boxes represent generic proteins, while peach-colored boxes represent viral proteins. Red-colored interactions are inhibitions, while the black interactions are activations. The diagram follows SBGN Process Description graphical notation guidelines for the standardized representation of biological mechanisms [50], making the diagram both human and machine-readable. The diagram is part of the COVID-19 Disease Map repository [47].

Spain, Netherlands, Germany, Italy, Great-Britain, USA, ...) launched in March 2020. It has united 267 scientists to build an open-access, interoperable, and computable repository of COVID-19 molecular mechanisms [47]. The community uses biocuration of scientific literature, text mining, and AI solutions to accelerate content building, along with popular pathway databases such as REACTOME [48] and WikiPathways [49]. Systems Biology standard notation schemes, like the Systems Biology Graphical Notation, SBGN [50], are used to represent the biological mechanisms. In these diagrams, nodes are biochemical entities, and arcs between nodes denote interactions between entities using standardized semantics depending on the form of the arc. An example is provided in Box 9. The Systems Biology Mark up Language (SBML) [51] is also used for modeling networks and creating content that is both human and machine-readable. This rich scientific ecosystem is described in Box 11.

In December 2020, the molecular events covered in the repository included 21 diagrams in MINERVA build, 18 diagrams in the Wikipathways collection, and 2 diagrams in REACTOME describing, among others, the virus replication cycle and subversion of host defenses, the virus attachment and entry, replication and release,... In addition to creating curated, standardized and interoperable diagrams, the community also develops tailor-made analytical and modeling pipelines for data integration and network analysis, and computational modeling.

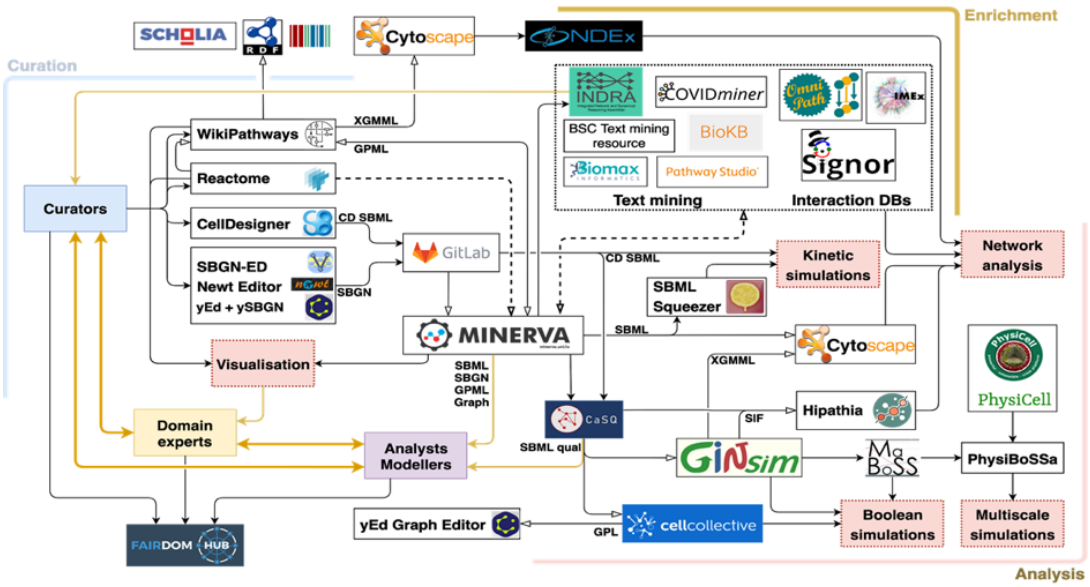
From static molecular maps to dynamic Boolean networks. Molecular maps are well-adapted to describe the pathways involved in a biological system, but as they are static, they cannot account for the dynamical behavior of the system. Studying dynamic behavior (how a system evolves over time

BOX 10. Example of a Boolean model



The inferred Boolean model corresponding to the BOX 9 diagram (left) and a real-time simulation example (right). Using the tool CaSQ [52], the Process Description Apoptosis diagram is now compressed in an Activity Flow-like graph, where only regulations and signaling flow are kept. Each interaction is described with a preliminary logical formula that defines the behavior of the nodes at each discrete time step, based on the topology of the graph and the semantics encoded. In the illustrated example, we can see a real-time simulation using the web platform Cell Collective [44]: when the death signal is activated via the FAS and FAS ligand complex (100 activity in the simulation graph), the Apoptotic phenotype will get activated after a few time steps (starts from 0 and increases to 100 activity)

BOX 11. Ecosystem of the COVID-19 Disease Map effort



The multidisciplinary, large-scale community effort of the COVID-19 Disease Map project develops interfaces between various platforms and tools. The community comprises Curators that use a variety of diagram editors and platforms, and benefit from the text mining and AI solutions developed or adapted within the community. Curators communicate with Analysts and Modellers, who use the curated content to create executable models of disease mechanisms. The community supports standardization efforts and the use of Systems Biology standards where possible. It promotes transparency and reproducibility by adhering to FAIR guiding principles (findable, accessible, interoperable, and reusable) and by using public repositories to share code and files. The figure here [47] illustrates the rich ecosystem of interoperable tools and platforms that the community members employ. It also shows the formats that facilitate the data exchange.

or in response to a given stimulus) requires an inherent execution scheme that explicitly describes the rules of the system's regulation. Boolean networks represent the simplest form of a qualitative model and are very often used to model gene regulation or signaling events. In a Boolean network, a node can be assigned the value 0 or 1 (0 for inactive or absent, 1 for active), and arcs between nodes can describe activation or inhibition. Boolean functions using logical operators (AND, OR, NOT) describe the rules that govern the regulation of each node based on the state of their regulators at every update of the system. The two most popular updating schemes are the synchronous—where all nodes of the network can be updated at the next step (deterministic behavior), and the asynchronous—where only one node of the network is allowed to change its state at the next step. Despite their simplistic nature, Boolean networks can capture most of the dynamics of a biological system by identifying steady states and complex attractors. These models are qualitative and can be used to address questions such as which phenotype will be activated in a given set of initial conditions [43].

The development of CaSQ has allowed the direct translation of static molecular maps to fully executable Boolean models with generalized logical rules based on the topology and semantics of the original molecular map [52]. The generated models allow for *in silico* simulations, perturbations, and predictions using various modeling platforms and tools such as Cell Collective [44], GINsim [53], and BoolNet [45], to name a few. The models are encoded in SBML qual format, a standard for qualitative, logic-based models in biology.

Computational modeling efforts. Besides dynamical models that account for time in the simulations, more mechanistic models of pathways also provide a way to study the impact of gene activity on phenotype. Among these, Hipathia is a tool that analyses transcriptomic and/or genomic data and calculates cellular profiles [56]. The diagrams of the COVID-19 Disease Map project have been implemented in the CoV-Hipathia version, where expression data are used to highlight upregulated or downregulated areas in the map. Hipathia can also derive pathway information from databases such as OmniPath [57] or SIGNOR 2.0 [58]. Another important community effort consists of the collaborative creation of a multiscale simulation model to study SARS-CoV-2 dynamics in lung tissue. The model was built rapidly and shared internationally as open-source code with an online interactive model that everybody can access and expand. The consortium behind this large-scale community effort includes experts across virology, immunology, mathematical biology, quantitative systems physiology, and high-performance computing [59]. An integrated host-virus genome-scale metabolic model of human alveolar macrophages and SARS-CoV-2 was also proposed [60]. The analysis of metabolic changes using flux balance analysis (FBA), a mathematical method for simulating metabolism in genome-scale reconstructions of metabolic networks, for both uninfected and infected host cells, demonstrated different profiles for host cells and the virus. Based on the hypothesis that modulations in the metabolic level could have different impacts on host and virus, the researchers studied the knock-out of the guanylate kinase (GK1, which decreased the virus's growth while leaving the host unaffected. Further in vitro testing is required to assess any potential therapeutic effect of GK1 inhibitors on SARS-CoV-2 infections.

Assessing the impact of environmental stressors. A computational systems biology approach was also employed to study whether environmental stressors, such as endocrine disrupting chemicals (EDCs), could contribute to certain chronic diseases and aggravate the course of COVID-19 [61]. The scientists compiled relevant datasets extracting biological associations of major EDCs to proteins from the CompTox database and COVID-19 comorbidities from the GeneCards and DisGeNET databases. A tripartite network (EDCs-proteins-diseases) was developed to identify proteins overlapping between the EDCs and the diseases. The Th17 and the AGE/RAGE signaling pathways were identified as possible targets of EDCs and as contributors to COVID-19 severity, thereby highlighting possible links between exposure to environmental chemicals and disease development.

Internet resources

- The COVID-19 Disease Map project: `https://covid.pages.uni.lu`

- The MINERVA build of COVID-19 Disease Map: https://covid19map.elixir-luxembourg.org/minerva
- The COVID-19 Pathways Portal on WikiPathways: https://www.wikipathways.org/index.php/Portal:COVID-19
- The REACTOME SARS-CoV-2 (COVID-19) infection pathway: https://reactome.org/content/detail/R-HSA-9694516
- The CoV-Hipathia: http://hipathia.babelomics.org/covid19
- The PhysiCell prototype for SARS-CoV-2: http://physicell.org/covid19
- The GeneCards website: https://www.genecards.org

Further reading on computational systems biology

[40] Computational systems biology. H. Kitano, *Nature*, 420, 206–210 (2002). doi.org/10.1038/nature01254

[41] Foundations in systems biology. Kitano, H, *MIT Press* (2001)

[42] Quantitative and logic modelling of molecular and gene networks. N. Le Novère, *Nat Rev Genetics.*, 16(3):146-58 (2015). doi:10.1038/nrg3885

[43] A practical guide to mechanistic systems modeling in biology using a logic-based approach. A. Niarakis and T. Helikar, *Briefings in Bioinformatics* (2020). doi.org/10.1093/bib/bbaa236

[44] The Cell Collective: Toward an open and collaborative approach to systems biology. T. Helikar, B. Kowal, S. McClenathan, S. et al. *BMC Systems Biolology*, 6, 96 (2012). doi.org/10.1186/1752-0509-6-96

[45] BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. C. Müssel, M. Hopfensitz, H.A Kestler, H.A., et al. *Bioinformatics* 26(10):1378–1380 (2010). doi.org/10.1093/bioinformatics/btq124

Bibliographical sources on computational systems biology

[46] COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. M. Ostaszewski, A. Mazein, M.E. Gillespie, et al. *Scientific Data* 7, 136 (2020). doi.org/10.1038/s41597-020-0477-8

[47] COVID-19 Disease Map, a computational knowledge repository of SARS-CoV-2 virus-host interaction mechanisms. M. Ostaszewski, A. Niarakis, A. Mazein et al. *bioRxiv* (2020). doi.org/10.1101/2020.10.26.356014

[48] The reactome pathway knowledgebase. B. Jassal, L. Matthews, G. Viteri, et al. *Nucleic Acids Research* 48(D1):D498-D503 (2020). doi:10.1093/nar/gkz1031

[49] WikiPathways: connecting communities. M. Martens, A. Ammar, A. Riutta, A., et al. *Nucleic Acids Research* 49(D1):D613–D621 (2021). doi.org/10.1093/nar/gkaa1024

[50] Le Novère, N., Hucka, M., Mi, H., et al. The Systems Biology Graphical Notation. *Nat Biotechnol.*, 2009 Aug;27(8):735-41. doi: 10.1038/nbt.1558

[51] SBML Level 3: an extensible format for the exchange and reuse of biological models. S.M. Keating, D. Waltemath, M. König, et al. *Molecular Systems Biology* 16:e9110 (2020. doi.org/10.15252/msb.20199110

[52] Automated inference of Boolean models from molecular interaction maps using CaSQ. S.S. Aghamiri, V. Singh, A. Naldi et al. *Bioinformatics* 36 (16): 4473–4482 (2020). doi.org/10.1093/bioinformatics/btaa484

[53] Logical modelling of regulatory networks with GINsim 2.3. A. Naldi, D. Berenguier, A. Fauré et al. *Biosystems* 97(2):134-139 (2009). doi.org/10.1016/j.biosystems.2009.04.008

[54] Human Coronavirus: Host-Pathogen Interaction. T.S. Fung and D.X Liu, *Annual Review of Microbiology* 73(1): 529-557 (2019)

[55] CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. A. Funahashi, N. Tanimura, M. Morohashi, et al. *BIOSILICO*, 1:159-162 (2003). doi:10.1016/S1478-5382(03)02370-9

[56] High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Hidalgo, M.R., Çubuk, C., Amadoz, A., *Oncotarget* 3(8), 5160–5178 (2017)

[57] OmniPath: guidelines and gateway for literature-curated signaling pathway resources. D. Turei, T. Korcsmaros and J. Saez-Rodriguez, *Nature Methods* 13(12) (2016)

[58] SIGNOR 2.0, the SIGnaling Network Open Resource 2.0. L. Licata, P. Lo Surdo, M. Iannuccelli, M., et al. *Nucleic Acids Research*, 8;48(D1):D504-D510 (2020). doi:10.1093/nar/gkz949

[59] Rapid community-driven development of a SARS-CoV-2 tissue simulator. M. Getz, Y. Wang, G. An, et al. *bioRxiv* (2020). doi.org/10.1101/2020.04.02.019075

[60] FBA reveals guanylate kinase as a potential target for antiviral therapies against SARS-CoV-2. A. Renz, L. Widerspick, A. Andreas Dräger, *Bioinformatics* 36(2): i813–i821 (2020). doi.org/10.1093/bioinformatics/btaa813

[61] Endocrine disrupting chemicals and COVID-19 relationships: A computational systems biology approach. Q. Wu, X. Coumoul, P. Grandjean, R. Barouki, et al. *Environment International*, 106232 (2020). doi.org/10.1016/j.envint.2020.106232

[62] The DisGeNET knowledge platform for disease genomics. J. Piñero, J.M. Ramírez-Anguita, J. Saüch-Pitarch et al. *Nucleic Acids Research* 48-D1, D845–D855 (2020). https://doi.org/10.1093/nar/gkz1021

2.3 Genetic susceptibility to the disease

As a complement to the analysis of the virus and environmental stressors, it is of great interest to understand the diversity of host responses. Indeed, one striking aspect of COVID-19 is how greatly the severity of its clinical manifestations varies across patients. While most people are asymptomatic or experience only mild symptoms, others will have a severe life-threatening response, which may be independent of their age or preexisting health condition(s). This type of observation suggests that genetic differences between individuals might contribute to different reactions to the disease. The hypothesis can be investigated by collecting genetic information from a large panel of healthy and infected individuals. Variations across individuals' genomes are then studied to identify possible links to the disease severity. Such studies on large population samples result in a better understanding of the infection susceptibility and lead to multiple potential applications: identifying patients at greatest risk, adapting the treatments, designing clinical trials, or even discovering new genetic targets against the virus. This is, however, a narrow road. The contribution of genetics to the susceptibility to, or severity of, a multifactorial disease such as COVID-19 may be quite small compared to other physiological and/or environmental factors. Moreover, the genetic contribution may result from many genomic positions, each contributing to a weak extent, rather than from a single mutation. In this context, the association signal may be very hard to detect. Studies therefore require panels of thousands of individuals in order to reach a reasonable statistical power (capacity of detection) as well as the use of advanced statistical models to prevent erroneous detection.

In June 2020, the first article identifying a genetic signal in the human genome was published [63]. The authors identified two regions, located on chromosome 3 and chromosome 9, through a statistical analysis called *GWAS* (Genome-Wide Association Studies. The analysis was conducted on genomic data collected at seven hospitals in the Italian and Spanish epicenters of the pandemic in Europe: 2,000 infected patients and 2,000 control participants. GWA studies collect two types of information for each individual on the panel: phenotypic information (e.g., infection level), and genotypic information. The DNA sequence of each individual is read at different positions over the genome, and their alleles (versions) are identified. Assuming that only two different alleles can be observed at a specific position, the whole

panel can be split into two subsamples – individuals having the first allele and individuals having the second one – allowing the disease severity measured in each subsample to be compared. Positions at which a significant difference between the two subsamples are observed are said to be "associated" to the disease severity.

Although the basic problem is quite simple, one needs to account for additional factors that may affect the trait and potentially blur the biological signal of interest. These factors include the gender or the age of the individuals, and the population structure of the panel. All these factors - along with the effect of the marker under study - are included in a regression model in order to quantify their respective impact on the disease severity. When members of the same family are included in the panel, the model becomes a *mixed* logistic regression model that includes a similarity matrix describing the levels of relatedness between all individuals. From a methodological point of view, running a GWA study requires evaluating the association between the trait of interest and each position along the genome. It is now standard practice to consider and query a very large number of positions (e.g., more than 8 million for this association study), each query requiring the fitting of the aforementioned logistic model, which results in a significant computational burden.

Interestingly, the identification of the genomic region on chromosome 9 can be related to previous clinical observations: individuals in blood group A were at higher risk of severe disease, while individuals in blood group O experienced a protective effect. This phenomenon might be explained by the fact that the genomic region in question carries the ABO genes that code for blood group. Regarding the region of interest on chromosome 3, further studies based on population genetics and evolutionary models have established that the specific form of the involved genomic segment could be inherited from Neanderthals, which creates an unexpected link between paleogenomics and COVID-19 [64]. This information has been widely reported in the mainstream press. This form occurs at a frequency of approximately 30% in South Asia, 8% in Europe, and more rarely in Africa, which could explain the initial geographical distribution of the disease.

During the following months, additional cohorts were recruited and analyzed. Like other data previously mentioned in this report, the results of each GWA study are usually available in public databanks, with full documentation and interoperable formats. Sharing individual genetic data raises ethical issues that should be addressed by ethics committees and data anonymization procedures. But the scientific value of this data is important. The availability of data is crucial for *meta-analysis*, which consists of performing a joint statistical analysis of all available results obtained from different studies in order to reach a higher detection level and identify new genetic markers, such as those involved in COVID-19. Meta-analyses require efficient and scalable algorithms to fit such models and data. In the last few years, the bioinformatics community has made some significant contributions for this purpose, including the METAL method [65] that was recently applied to the joint analysis of GWAS for the COVID-19 disease. Currently, a full GWA study involving 100,000 individuals genotyped at tens of millions of positions can be performed in a matter of hours – or minutes, depending on the complexity of the fitted model [66, 67]. These methods even open the way to *global genetics*. For example, the American private company AncestryDNA, which specializes in genealogy and at-home genetic testing, conducted a large-scale analysis in Spring 2020 [68]. It collected over 500,000 COVID-19 survey responses between April and May 2020 with accompanying genetic data from its own database. This work confirmed the previously identified regions on chromosomes 3 and 9.

Internet resources

- The *COVID-19 Host Genetics Initiative*, hosted at `www.covid19hg.org`, gathers the GWAS results for meta-analysis reusability by the scientific community.

Bibliographical sources on the genetic susceptibility of SARS-CoV-2

[63] Genomewide association study of severe Covid-19 with respiratory failure. Severe Covid-19 GWAS Group, *New England Journal of Medicine* 383(16): 1522-1534 (2020). https://www.nejm.org/doi/10.1056/NEJMoa2020283

[64] The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. H. Zeberg and S. Pääbo, *Nature* 587, 610–612 (2020). doi.org/10.1038/s41586-020-2818-3

[65] METAL: fast and efficient meta-analysis of genomewide association scans. C.J. Willer, Y. Li and G.R. Abecasis, *Bioinformatics*, 26(17):2190-2191 (2010). doi.org/10.1093/bioinformatics/btq340

[66] A resource-efficient tool for mixed model association analysis of large-scale data. L. Jiang, Z. Zheng, .T. Qi, K.E. Kemper, N.R. Wray, P.M Visscher and & Yang, J. Yang (2019), *Nature Genetics* 51:1749–1755 (2019). doi.org/10.1038/s41588-019-0530-8

[67] Mixed-model association for biobank-scale datasets. P.R. Loh, G. Kichaev, S. Gazal, A.P. Schoech and A.L. Price, *Nature Genetics* 50(7), 906-908 (2018). doi.org/10.1038/s41588-018-0144-6

[68] AncestryDNA COVID-19 Host Genetic Study Identifies Three Novel Loci. H.G. Roberts et al. *medRvix* (2020). doi.org/10.1101/2020.10.06.20205864

3 Understanding the past, anticipating the future: the origins and dynamics of SARS-CoV-2 evolution

Determining the origins of the SARS-CoV-2 virus is crucial for multiple reasons. Beyond its scientific relevance, precisely reconstructing the chain of events that led to the COVID-19 pandemic would most likely have a strong impact on geopolitics. Ecologists have also suggested that there could be a causal relationship between the current erosion of biodiversity due to human activities and the advent of pandemics[1]. Deciphering the origins of SARS-CoV-2 would provide crucial evidence for or against this stance. But most importantly, knowing the exact cause of the pandemic could help prevent similar crises in the future. We offer below a non-exhaustive overview of the available evidence about the origins of SARS-CoV-2.

3.1 Hypothesis on the origins of the virus: an overview

After the first public database searches described in Section 1.2, sequence alignments targeting betacoronavirus genomes and some functional regions of these genomes provided a more precise picture of the origin of SARS-CoV-2.

First, they indicated that the strain most closely related to SARS-CoV-2 was collected in 2012/2013 inside a cave in the Yunnan province (China). This strain, named RaTG13 [69, 70], has a particularly interesting story to tell. First of all, the corresponding sample was obtained after six miners who worked in this cave fell ill, showing symptoms similar to those displayed by patients suffering from COVID-19. Also, a large population of bats is found in this particular cave. The sample of interest was retrieved by a research lab specializing in viral diseases transmitted to humans by these animals. The sample was then transferred to the city of Wuhan, where the lab in question is located, not far from the food market that is suspected to be the origin of the pandemic.

However, RaTG13 and SARS-CoV-2 genomes are only ∼96% identical, and the most recent common ancestor of these two viruses is at least 20 years old [71]. Although these two strains are clearly closely related, RaTG13 could only be at the origin of the current pandemic if that strain had already been circulating and evolving in humans (or a closely related host) for a few years without triggering any public health/sanitary alert. Furthermore, RaTG13 is unable to infect human cells [72], making it an unlikely candidate for the original strain. Moreover, in several previous epidemics caused by coronaviruses, bats did not directly transmit the virus to humans. For instance, a civet is thought to be at the origin of the 2003 epidemic provoked by the SARS-CoV virus. In 2012, a coronavirus transmitted by camels to humans was responsible for the so-called Middle East respiratory syndrome, MERS. There are also other examples where bats were ruled out as the last known host before diffusion in the human population. However, in all these cases, the chiropterans are thought to act as the natural reservoir for the coronavirus strains associated with these zoonotic diseases.

[1] see for instance https://ipbes.net/pandemics

The role of recombinations. The existence of an intermediate host that could have facilitated the transmission of SARS-CoV-2 from bats to humans therefore remains an open question. Pangolins are known for being infected by viral strains very close to SARS-CoV-2. In fact, the Spike protein, which plays a central role in the entry of the virus in human cells, displays a very similar sequence in the coronaviruses circulating amongst pangolins and in SARS-CoV-2. The identity percentage within the receptor-binding domain (RBD) of this protein is 97.4%, while the same genomic region in RaTG13 is only 83.3% identical to SARS-CoV-2 [73, 74]. However, other viral strains circulating in bats show insertions that encode for residues essential to entering lung cells and other tissues in humans [74]. These elements suggest that the intervention of an intermediate host, pangolin or other, is not an absolute requirement for explaining the origins of the pandemic, and that *recombination* between multiple strains of related coronaviruses could have played a central role in the process. Besides mutations, recombination is another natural way viruses evolve. When a cell is infected by several strains of the same virus, their genomes can mix and produce a novel strain. Here, co-infection of bats with multiple strains, thereby potentially combining genomes from distinct strains, has been documented in the recent past [69]. Further investigations about the origins of the SARS-CoV-2 will therefore probably concentrate on the mechanisms that assembled the mosaic genome found in the SARS-CoV-2 virion that infected the first human patient.

The role of adaptation. Alongside recombination, adaptation is of the utmost importance in explaining how SARS-CoV-2 successfully disseminated in the human population on a global scale. Adaptive processes are likely to have shaped the evolution of the RBD region of the spike protein so as to optimize its binding to the ACE2 protein, the human enzyme that lets the virus enter the cells (see Section 2.1). It is not clear, however, whether adaptation took place after or before the zoonotic event [75]. Furthermore, as opposed to SARS-CoV, which showed signs of rapid genomic adaptation to humans at the start of the epidemic, SARS-CoV-2 has followed a different evolutionary trajectory [76]. Apart from a few notable exceptions (see below), SARS-CoV-2 seems to have been evolving according to a neutral regime [77], at least for most of 2020. This behavior is in line with the hypothesis that the virus was already well adapted to humans at the start of the pandemic. A slightly modified form of SARS-CoV-2 could have thus circulated among humans for some time. Quiet circulation amongst other mammals is another possibility. In fact, the human ACE2 receptor is very similar to that found in several domestic and/or laboratory animals and livestock such as hamsters, cows, goats or sheep [78].

It is worth noting, however, that, despite an overall neutral evolution, adaptation is likely to govern the fate of some mutations of the SARS-CoV-2 genome. Hence, the replacement of an aspartate by a glycine residue at position 614 of the spike protein (noted D614G) has seen its frequency and geographic range increase rapidly during spring 2020. This particular mutation is associated with higher viral loads and is over-represented among younger age cohorts [71, 79]. Another variant, first detected in the UK, also sharply increased in frequency late in 2020. This lineage, termed B.1.1.7 (or 501Y.V1), sustained an unusually large number of substitutions. Four out of the 14 B.1.1.7-specific mutations are located, once again, in the spike protein, including, most notably, the replacement of an Asparagine residue at position 501 (see Box 8) by a Tyrosine in the RBD region. Evidence indicates that these changes may facilitate the transmission of the virus, suggesting that natural selection played a role in the processes that led to a worldwide increase in the frequency of this lineage [80, 81].

Questions related to the actual adaptation processes taking place are nonetheless difficult to answer given the relatively low amount of genetic diversity observed amongst circulating SARS-CoV-2 viruses to date. Also, efforts to clarify the origins of the virus would greatly benefit from new samples from different places on earth, particularly China. The cave in the Yunnan province where RaTG13 was found is undoubtedly one the most interesting areas to search. Re-analyzing the tissues collected from some of the Yunnan miners who fell sick in 2012/2013 would also be of utmost interest. It would furthermore be relevant to search for viruses related to SARS-CoV-2 in data banks of human tissues that are conserved by many research laboratories around the world.

Bibliographical sources on the origins of the virus

[69] Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. X.-Y. Ge, N. Wang, W. Zhang, B. Hu, B. Li, Y.-Z. Zhang, J.-H. Zhou, C.-M. Luo, X.-L. Yang, L.-J. Wu, et al. *Virologica Sinica*, 31(1):31–40, 2016.

[70] A pneumonia outbreak associated with a new coronavirus of probable bat origin. P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al. *Nature*, 579(7798): 270–273, 2020b

[71] Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. M. F. Boni, P. Lemey, X. Jiang, T. T.-Y. Lam, B. Perry, T. Castoe, A. Rambaut, and D. L. Robertson. *Nature Microbiology*, 5:1408–1417 (2020)

[72] SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. A. G. Wrobel, D. J. Benton, P. Xu, C. Roustan, S. R. Martin, P. B. Rosenthal, J. J. Skehel, and S. J. Gamblin. *Nature structural & molecular biology*, 27(8):763–767 (2020)

[73] Identifying SARS-CoV-2-related coronaviruses in Malayan Pangolins. T. T.-Y. Lam, N. Jia, Y.-W. Zhang, M. H.-H. Shum, J.-F. Jiang, H.-C. Zhu, Y.-G. Tong, Y.-X. Shi, X.-B. Ni, Y.-S. Liao, et al. *Nature*:1–4 (2020)

[74] A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. H. Zhou, X. Chen, T. Hu, J. Li, H. Song, Y. Liu, P. Wang, D. Liu, J. Yang, E. C. Holmes, et al. *Current Biology*, 2020a

[75] The proximal origin of SARS-CoV-2. Andersen, Kristian G and Rambaut, Andrew and Lipkin, W Ian and Holmes, Edward C and Garry, Robert F. *Nature medicine 26* (2020)

[76] SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence? S. H. Zhan, B. E. Deverman, and Y. A. Chan. *bioRxiv* (2020)

[77] No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. L. van Dorp, D. Richard, C. C. Tan, L. P. Shaw, M. Acman, and F. Balloux. *bioRxiv* (2020)

[78] Tracing the origins of SARS-COV-2 in coronavirus phylogenies. E. Sallard, J. Halloy, D. Casane, E. Decroly, and J. van Helden. *médecine/sciences*, 36(8-9) (2020)

[79] Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. B. Korber, W. M. Fischer, S. Gnanakaran et al. *Cell*, 182(4):812–827 (2020)

[80] Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. A. Rambaut, N. Loman, O. Pybus, W. Barclay, J. Barrett, A. Carabelli, T. Connor, T. Peacock, D. L. Robertson, and E. Volz. *virological.org* (2020). https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563

[81] Transmission of SARS-CoV-2 lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. E. Volz, S. Mishra, M. Chand et al., *Report from MRC Centre for Global Infectious Disease Analysis*, (2020). https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-12-31-COVID19-Report-42-Preprint-VOC.pdf.

3.2 Using phylogenetics to reconstruct and monitor the pandemic

Comparing and aligning genetic sequences conveys a wealth of information about the evolutionary relationships between genes, genomes, populations, and species, depending on the scale of the analysis. It is possible to depict these relationships by creating a *phylogenetic tree* that explains the evolutionary history of the sequences. The phylogenetic tree represents the backbone along which genetic mutations take place. In practice, multiple sequence alignments of DNA and/or protein sequences are fed as input for phylogenetic tree-building methods. The tree reconstruction techniques then rely on a sound mathematical criterion to reconstruct the tree that best fits the available data.

BOX 12. How to construct phylogenetic trees: likelihood-based techniques

The *likelihood criterion* is a very well-characterized statistical criterion that is commonly used to infer phylogenies. The likelihood of a phylogenetic model corresponds to the probability of the observed molecular data, i.e., the multiple sequence alignment, given the model. In other words, the phylogenetic model can be considered here as a "stochastic generator" that mimics molecular evolution by randomly mutating sequences that evolve along its edges. The sequences observed at the tips of the tree are the product of this stochastic simulation. Our goal is to find the parameters of the generator that best fit the observations. Evaluating the likelihood of a phylogeny relies on a recursive algorithm that corresponds to a depth-first post-order tree traversal. The core of the recursion consists of evaluating the probability of the subtree below a particular node given the ancestral sequences that could be observed at these nodes. Because the different positions along the analyzed gene or genome are assumed to be independent and to evolve according to the same model (i.e., the columns in the alignment are all independent and identically distributed), the likelihood score can be expressed as the product of the likelihoods evaluated at every sequence position. In layman's terms, using the likelihood score as a guide for the inference guarantees the best exploitation of the data with respect to the parameters of interest, i.e., the phylogenetic tree in the present context. It should thus not come as a surprise that, in practice, likelihood-based tree reconstruction techniques provide, on average, the most accurate phylogenetic tree estimates, assuming that the probabilistic model describing the mutation process is not too distinct from the truth. Two options are then available: one can either (1) search for the parameter values that maximize the likelihood function, or (2) sample these values to a frequency proportional to their likelihoods or *a posteriori* probabilities. Maximizing the likelihood of a phylogenetic model relies on heuristic algorithms that alternate between updating the tree topology, i.e., the structure of the graph, and adjusting the continuous parameters of the model (the edge lengths, the relative rates of different types of substitutions, e.g., transitions vs. transversions). Similar operators are used by techniques that sample the model parameters instead of optimizing them. A major difference with the maximization approach is that sampling techniques sometimes retain sub-optimal solutions, and the set of all sampled solutions defines the whole posterior distribution of parameters instead of point estimates. Hence, sampling and maximization techniques do not solve the same problem, thereby explaining why sampling techniques are inherently slower than maximization approaches. Note, however, that quantifying the variability of parameter estimates in the context of maximum likelihood generally requires running non-parametric bootstrap analyses, which can make them just as computationally costly as sampling approaches. Evaluating the likelihood score is the main computational bottleneck of these inference techniques. Yet, important progress has been made in this area over the last fifteen years. Research has focused here on re-using partial likelihood scores, i.e., likelihoods of subtrees that are common to multiple sites in the alignment or likelihoods of subtrees not affected by operators that update the phylogeny when exploring new solutions [82, 83]. Implementing algorithms that accommodate for the architecture of modern computing units such as GPU has also helped speed up the calculation [84].

The limits of likelihood methods. Phylogenetic reconstruction is a central field of bioinformatics. Like many other areas of this discipline, a large array of algorithms have been proposed over the years to cope with data of increasing size and complexity. From a historical perspective, maximum-likelihood is the method of choice. Introduced by Felsenstein in 1981, the principle is to find the tree and the parameters of the nucleotide substitution model that maximize the probability of the multiple sequence alignment. The maximum-likelihood paradigm rests on sound mathematical properties, but it is computationally intensive. It has been the subject of many fruitful subsequent developments to speed up the calculations. We present the highlights in Box 12. Nonetheless, despite all these optimizations and heuristics, likelihood-based techniques start to struggle when processing data sets with more than ∼1,000 distinct sequences. Hence, maximum-likelihood was used, for example, to construct the reference phylogeny of March 2020, which relies on 176 SARS-CoV-2 genomes. The tree is shown in Box 13. However, the rapid accrual of SARS-CoV-2 genomes in the GISAID database made it impossible to carry on using maximum-likelihood, or Bayesian sampling techniques for that matter. Faster, but less accurate techniques, which rely on the *minimum evolution principle*, are now used to update the phylogeny of SARS-CoV-2.

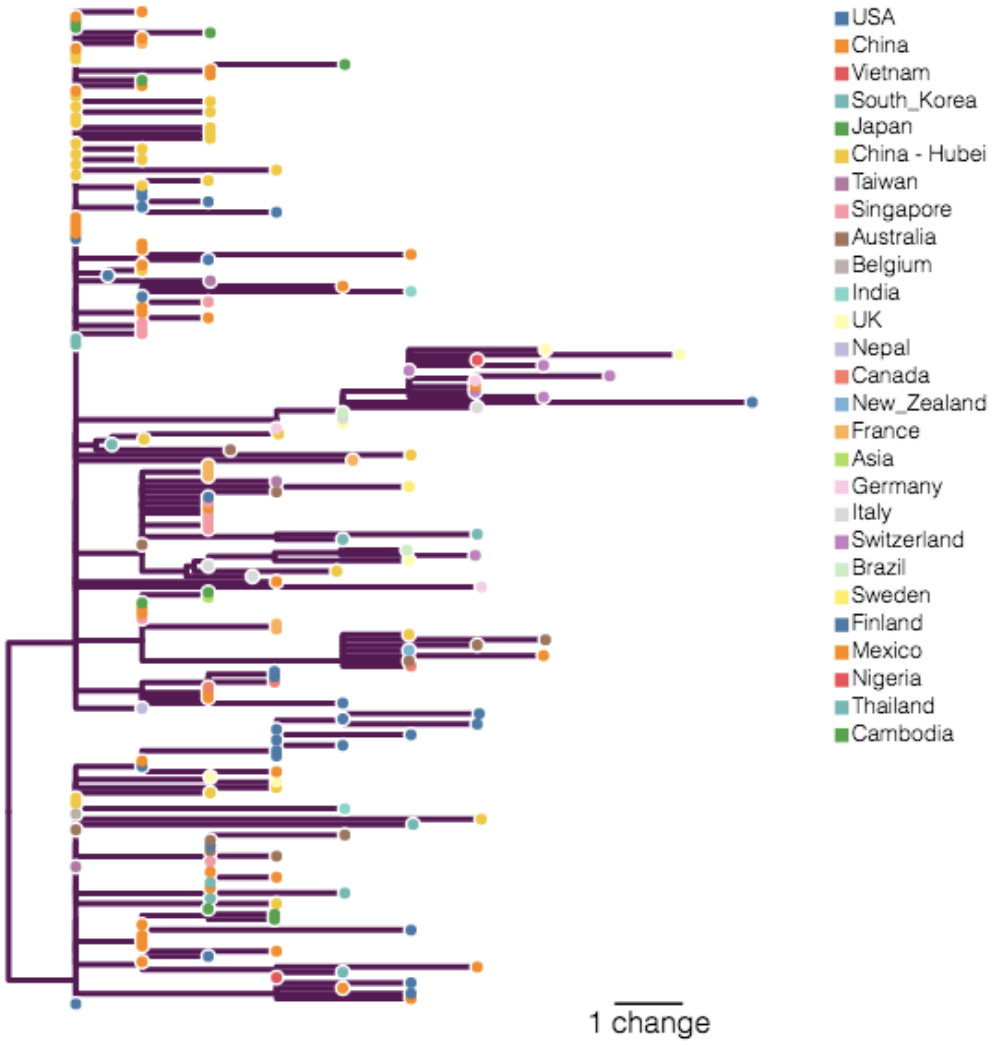
Reference SARS-CoV-2 phylogenies with distance methods. The minimum evolution criterion applies to the matrix of pairwise distances between sequences instead of the raw data (i.e., the multiple sequence alignment). Assuming that such a matrix is already available (it is usually estimated using the same stochastic models of evolution as those used by likelihood-based inference approaches), the rationale here is to choose, for a given tree topology, edge lengths that best fit the estimated pairwise distances. A weighted least-squares criterion that accommodates for the variance of distance estimates is used here to work out the lengths of all edges in the tree. The selected tree topology minimizes the sum of edge lengths estimated in this way, thereby following the principle of Occam's razor.

The pitfall of recombinations. Until now, phylogenetics software tools have been able to handle the large number of SARS-CoV-2 genomes produced daily. There is, however, a main theoretical frontier with these existing methods: neither likelihood- nor distance-based approaches account for recombinations, whereas we have seen that this process is central to understanding the origins of the COVID-19 pandemic (see Section 3.1). As of January 2021, recombination does not seem to have substantially impacted the dynamics of the pandemic [85]. Yet, this situation is likely to change with an increasing amount of observable polymorphism as mutations keep accumulating on the SARS-CoV-2 genome. Moreover, evidence indicates that recombination played a role in the dynamics of previous CoV epidemics [86]. Because recombination exchanges genetic material from distinct lineages, it greatly complicates the phylogenetic signal conveyed by whole genomes, with different portions of these genomes having distinct evolutionary histories. Properly accommodating for this feature of molecular evolution is therefore important. Detecting the presence of recombination-driven breakpoints within an alignment of homologous sequences is the first step towards tackling this issue. Phylogenetic models that explicitly account for recombination are also being developed. Sophisticated methods that combine graph algorithms and probabilistic modeling are involved here. Further developments are required here in order to alleviate biases in substitution rate and node age estimates due to recombination.

Internet resources

- `nextstrain.org` is specialized in monitoring pandemics and epidemics using graphical tools based on phylogenetics. It is arguably one of the most reliable sources of information regarding the evolution of SARS-CoV-2 (and other viruses). The reference phylogeny for SARS-CoV-2 is *de facto* that proposed on this site. The tree is built using the software tool FastTree [87], which implements a hybrid method inspired by the same agglomerative algorithms that distance-based methods use. The site also provides detailed reports on the dynamics of the pandemic in different countries/regions, with the marriage of phylogenetics and geography. Indeed, phylogenetics, when applied to geo-referenced genetic data, can help determine the spatial location of the ancestor of a sample of genomes, thereby providing helpful information about the location where a pandemic

BOX 13. Phylogenetics analysis from 176 early SARS-CoV-2 genomes



The analysis uses 176 full-length SARS-CoV-2 genomes that were available on the GISAID and NCBI platforms before March 2020: sequences were collected from December 2019-February 2020. Each color corresponds to the country where the genome was sampled. The length of the branch is proportional to the number of mutations (changes). This shows that, at that time, there was limited genetic variation in the sampled viruses. This lack of diversity is indicative of a relatively recent common ancestor for all these viruses. The tree was built with the PhyML software, which implements a very efficient maximum-likelihood-based algorithm.

Source: virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356.

originated and the pace at which it diffuses across space. Displaying spatial information at the tips (the leaves) of the tree by coloring the corresponding sequences with their sampling locations, for instance, is also tremendously helpful for deciphering the migration patterns underlying a pandemic. One can, for example, quickly determine whether viral strains circulating in a given country all descend from a single introduction event or from a succession of multiple introductions. Moreover, these specific events can be dated, provided that the rate at which mutations accumulate is known *a priori*, whereby "molecular times" (i.e., expected numbers of mutations) are translated into "calendar times". By combining phylogenetics and spatial information about the sampled genetic sequences, elucidating the geographic origins of most viral clusters has become feasible, even by non-specialists. This capability may explain the site's rapid popularity gain at the end of 2019-early in 2020. Note, however, that this tool does not display information about the uncertainty around phylogenetic model parameter estimates, including ancestral geographical locations of the virus. Given the low amount of genetic diversity displayed by SARS-CoV-2 sequences, combined with potential ambiguities regarding the geographic origin of a given virion (a given individual infected in China could have been "sampled" in Europe, for instance), ignoring this uncertainty can have serious consequences.

- `cov-lineages.org` has also been widely used for determining the phylogenetic placement of new SARS-CoV-2 strains. This approach relies on a backbone tree that comprises ∼30000 tips, which serves as a basis to determine the phylogenetic position of a few new sequences. Since the backbone tree is fixed, placing a new sequence within this tree is relatively fast, making the resource an efficient diagnostic tool (e.g., determining the country of origin of a given strain is quick and straightforward here).

Bibliographical sources on the use of phylogenetics to monitor the pandemic

[82] RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. A. Stamatakis, *Bioinformatics*, 22:2688–2690 (2006)

[83] New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. *Systematic Biology*, 59(3):307–321 (2010)

[84] BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. D. L. Ayres, M. P. Cummings, G. Baele, A. E. Darling, P. O. Lewis, D. L. Swofford, J. P. Huelsenbeck, P. Lemey, A. Rambaut, and M. A. Suchard. *Systematic Biology*, 68(6):1052–1061, (2019)

[85] No detectable signal for ongoing genetic recombination in SARS-CoV-2. D. Richard, C. J. Owen, L. van Dorp, and F. Balloux. *bioRxiv* (2020)

[86] Molecular evolution of human coronavirus genomes. D. Forni, R. Cagliani, M. Clerici, and M. Sironi. *Trends in microbiology*, 25(1):35–48 (2017)

[87] Fasttree 2–approximately maximum-likelihood trees for large alignments. M. N. Price, P. S. Dehal, and A. P. Arkin. *PloS one*, 5(3):e9490 (2010)

3.3 Tree generating models

Phylogenies convey much more information than the "simple" depiction of evolutionary relationships between organisms. Inferred trees may indeed serve as data for inferring parameters that characterize the demographic and epidemiological dynamics of a population. One can, for instance, estimate the so-called *effective population size* from the ages of the internal nodes of the tree. The rationale behind the concept of effective population size is relatively simple: the larger the viral population is, the further into the past one has to travel to find common ancestors of viruses sampled in more recent times. This idea is at the core of Kingman's coalescent. This probabilistic model was later generalized to accommodate for population sizes that may fluctuate over the course of evolution (in a deterministic fashion). It is

now commonplace to apply a coalescent model assuming an exponentially growing population size and test whether its fit to the phylogeny is better than that provided by a constant-size coalescent model.

This type of approach was used at the end of 2020 to demonstrate that the growth of the British variant B.1.1.7 population size in the UK was indeed exponential [81]. The exponential rate parameter can easily be translated into an amount of time required for the population size to double, which is an easy-to-interpret statistic that helps authorities evaluate the seriousness of a pandemic. Statistical modeling based on phylogenetic data thus provides a relevant alternative to the standard approach, which aims to infer a virus's prevalence in a population. Indeed, compared to counting methods, the coalescent is less sensitive to spatial and/or temporal variations of the sampling intensity. The coalescent, however, is only relevant over time scales that authorize sufficient genetic polymorphism so that the node ages in the phylogeny can be estimated accurately.

Other probabilistic tree-generating models, where lineages arise and die at rates that can be estimated from phylogenies, are also relevant from an epidemiology point of view. In the context of virus phylogenetics, these models assume that each branching event corresponds to a transmission event. The end of a lineage (i.e., a tip in the tree) corresponds either to the infected individual's death or recovery (in one case or the other, the corresponding viral lineage disappears). Phylogenies of viruses are therefore conceptually fairly remote from "standard" phylogenies where splits of lineages correspond to speciation events and terminal nodes to extinction (or sampling) events [88]. Accurately estimating these birth and death parameters when analyzing virus sequences is crucial since the ratio between the rate at which lineages split and die is an estimate of the reproductive number [89]. The death parameter itself is closely linked to the effective infection duration as it governs the expected duration of a lineage. For instance, by analyzing about 200 sequences collected in the GISAID database, Danesh et al. (2020) [90] were able to estimate that the median effective infection duration in France early in Spring 2020 was close to five days.

Tree-generating models can therefore help extract relevant information about the dynamics of a pandemic by analyzing genetic data from evolutionary trees. Nevertheless, the models presented above are far from perfect when considered through the lens of epidemiology. For instance, the proportion of susceptible individuals, i.e., the fraction of people that can potentially be infected, decreases as the pandemic escalates [91, 92]. Yet, for the sake of mathematical simplicity, both Kingman's coalescent and the standard birth-death tree models ignore this information. Phylodynamics [93] is a new and active research area that aims to circumvent these limitations and incorporate elements of epidemiology into modern population genetics models, all of which rely on building evolutionary trees.

The standard models in epidemiology, stochastic or deterministic, rely on incidence data, i.e., the number of new cases per time unit, to monitor the dynamics of an epidemic. As opposed to phylodynamics, these models assume that all observations are independent of one another, which may be problematic. Yet, this simplification authorizes the deployment of realistic models where the whole population may be structured both spatially and according to compartments (typically, susceptible, infectious, and recovered). This level of sophistication may be incorporated into the phylodynamics framework at the price of a high level of sophistication for evaluating the likelihood function [94, 95] (although see [96] for an attempt to tackle this issue). Moreover, collecting incidence (or prevalence) data remains quicker and less expensive than sampling the corresponding genetic sequences, and it is not always obvious why complex phylodynamics analyses should be performed in situations where straightforward epidemiological modeling suffices. Nonetheless, quantifying incidence strongly depends on the sampling intensity, i.e., detected cases represent a fraction of actual cases, and this fraction may depend on multiple (time and space-dependent) factors. Hence, both approaches have their strengths and weaknesses, and ongoing research efforts are required to make the most sensible use of the data available.

Bibliographical sources on the use of tree-generating models to monitor the pandemic

[88] S. Alizon. Phylogénies d'infections et phylodynamique. In G. Didier and S. Guindon, editors, *Comprendre l'évolution, approches mathématiques et informatiques*. ISTE-Wiley, (2020) https://hal.archives-ouvertes.fr/hal-02884408.

[89] On incomplete sampling under birth-death models and connections to the sampling-based coalescent. T. Stadler. *J Theor Biol*, 261(1):58–66 (2009)

[90] Early phylodynamics analysis of the COVID-19 epidemic in France. G. Danesh, B. Elie, Y. Michalakis, M. T. Sofonea, A. Bal, S. Behillil, G. Destras, D. Boutolleau, S. Burrel, A.-G. Marcelin, J.-C. Plantier, V. Thibault, E. Simon-Loriere, S. v. der Werf, B. Lina, L. Josset, V. Enouf, and S. Alizon. *medRxiv*, 2020.06.03.20119925, ver. 3 peer-reviewed and recommended by *PCI in Evolutionary Biology*, 2020.

[91] Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond. *Proceedings of the National Academy of Sciences*, 110(1):228–233 (2013)

[92] Bayesian phylodynamic inference with complex models. E. M. Volz and I. Siveroni. *PLOS Computational Biology*, 14(11):e1006546 (2018) doi.org/10.1371/journal.pcbi.1006546

[93] Unifying the epidemiological and evolutionary dynamics of pathogens. B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes, *Science*, 303(5656):327–32 (2004)

[94] Using an epidemiological model for phylogenetic inference reveals density dependence in hiv transmission. G. E. Leventhal, H. F. Günthard, S. Bonhoeffer, and T. Stadler, *Molecular biology and evolution*, 31(1): 6–17 (2014)

[95] Phylodynamic inference for structured epidemiological models. D. A. Rasmussen, E. M. Volz, and K. Koelle, *PLoS Computational Biology*, 10(4):e1003570 (2014)

[96] Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. E. Saulnier, O. Gascuel, and S. Alizon. *PLOS Computational Biology*, 13(3):e1005416 (2017)

A few words of conclusion

Scientific research on COVID-19 was conducted at an unusually fast pace in 2020; these efforts were indispensable for understanding and controlling the pandemic. In this race against time, bioinformatics has played an essential role, hand in hand with biology and medicine. We have presented in this document a brief tour of the state-of-the-art, illustrated by a selection of computational tools and methods that contributed to the understanding of SARS-CoV-2. This survey is not intended to be exhaustive, and our knowledge about SARS-CoV-2 is growing every day. For example, we did not mention third generation sequencing, single-cell sequencing or the formation of RNA secondary structures. Nevertheless, this report shows the maturity and the diversity of computational biology.

A point worth mentioning is that the bioinformatics community has a pioneering experience in open science. This concerns the development of open source software with high technology readiness level. This also concerns data collection, with numerous initiatives deployed for the development of open data access and the definition of universal formats for data storing and sharing following the FAIR principles (findability, accessibility, interoperability, and reusability). This intellectual and philosophical position has played a crucial role in accelerating the research since the beginning of the crisis.

Our overview also shows the integrated nature of COVID-19 research, and the coherence of bioinformatics sub-disciplines. Computational methods for sequence analysis are key to establishing and deciphering the genome of the virus. Phylogenetic studies play a central role in assessing the animal reservoir of pathogens and tracking mutations. Understanding genetic variations prompts questions on the structural biology side. In turn, a full qualification of interactions at the molecular level helps to refine the parameterization of interactions at the systems biology level, with a direct bearing on medicine. A matrix-like research effort (in the individual disciplines, but also in terms of integration) must therefore be pursued to improve therapeutics and medical protocols, but also to develop insights into zoonosis. We understand today that the fight against COVID-19 is a long journey. Long-term research efforts are still required to build a more robust and detailed picture of this episode. Scientific innovation and knowledge-building will be critical to help us face future challenges of the same sort.

Authors

- Samuel Alizon, DR CNRS, evolutionary ecology
  MIVEGEC, Univ. Montpellier, IRD, CNRS, Montpellier, France
- Frédéric Cazals, DR Inria, structural bioinformatics
  Université Côte d'Azur, Inria, France
- Stéphane Guindon, CR CNRS, phylogeny and evolution
  Department of Computer Science, LIRMM, CNRS and Université de Montpellier, Montpellier, France
- Claire Lemaitre, CR Inria, sequencing and genome analysis
  Univ Rennes, Inria, CNRS, IRISA, Rennes, France
- Tristan Mary-Huard, CR INRAE, statistical methods for quantitative genetics
  Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, Gif-sur-Yvette, France
  MIA, INRAE, AgroParisTech, Université Paris-Saclay, 75005, Paris, France
- Anna Niarakis, associate professor at Evry University, systems biology
  Univ Evry, University of Paris Saclay and Inria Saclay)
- Mikaël Salson, associate professor at University of Lille, sequencing and genome analysis
  Université de Lille, CNRS, Centrale Lille, CRIStAL UMR 9189, F-59000 Lille, France
- Céline Scornavacca, DR CNRS, phylogeny and evolution
  Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier, France
- Hélène Touzet, DR CNRS, sequencing and genome analysis (coordination of the report)
  Université de Lille, CNRS, Centrale Lille, CRIStAL UMR 9189, F-59000 Lille, France

This report is a production of the GDR Bioinformatique Moléculaire from CNRS: gdr-bim.cnrs.fr. The first date of publication is March 15, 2021.

The authors thank Rayan Chikhi and Pierre Peterlongo for fruiful comments on this document.

Comments and feedback are welcome. You can contact the authors at bioinfcov@univ-lille.fr.