ACS Medicinal Chemistry Letters

pubs.acs.org/acsmedchemlett

Viewpoint

How a Blockchain Approach Can Improve Data Reliability in the COVID-19 Pandemic

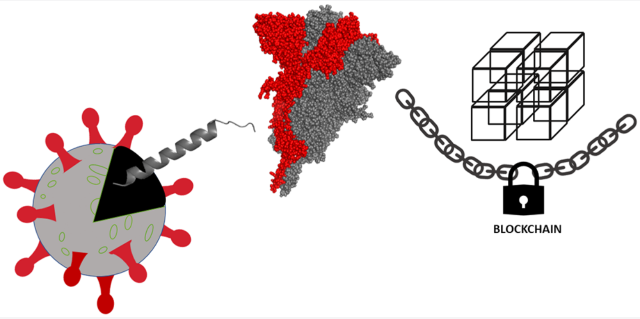
Pietro Cozzini,* Federica Agosta, Greta Dolcetti, and Gianfranco Righi

Cite This: ACS Med. Chem. Lett. 2022, 13, 517−519

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The rapid spread of COVID-19 made it necessary to quickly collect and share viral genomic sequences, sometimes making quantity prevail over the quality of information. Can research pay this price? Blockchain technology, based on the concept of a ledger that guarantees the authenticity and traceability of information, could be the best applicable solution.



KEYWORDS: *Blockchain, databases, data quality, COVID-19*

The global pandemic situation is widely well-known and changes every day. The virus changes to adapt its structure to improve the survival probability. Obviously, the mutations are random, and the survivor is the mutated virus whose Spike glycoprotein is able to establish stronger interactions with the human ACE2 protein and/or weaker interactions with antibodies. Thus, it is of paramount importance to know the structure of all the virus mutations found to assess better strategies to discover new Spike inhibitors or to plan new specific vaccine versions.

It is like a cops and robbers game where the virus is the robber, and the researchers are the cops. To have success in this battle, the cops need to be one step ahead; thus, it is important to be able to predict the most probable mutations we expect in the future. A huge collection of protein sequences is needed to pursue this goal.

Based on genomic sequences collected in recent months around the world, researchers, institutions, and organizations provide updated and diversified information relating to the pandemic spread. The National Center for Biotechnology Information, NCBI (https://www.ncbi.nlm.nih.gov/sars-cov-2/), for example, provides viral genome and protein and nucleotide sequences and structures. The WHO coronavirus (COVID-19) Dashboard (https://covid19.who.int/), instead, describes the global trend of infections by WHO Regions and summarizes measures taken by countries to limit the spread of COVID-19, while the European Centre for Disease Prevention and Control, ECDC (https://www.ecdc.europa.eu/en/covid-19/data), provides daily and weekly data based on the sequences available on GISAID, the world's largest repository of SARS-CoV-2 sequences.

The GISAID Initiative (https://www.gisaid.org/) was created to promote the sharing of data on avian influenza and then on the H1N1 pandemic in 2009 and the H7N9 epidemic in 2013. On January 10, 2020, it made available the first complete genome sequence of SARS-CoV-2, and since then it has become the main genomic sequence database in the world.

> *"The GISAID Initiative promotes the rapid sharing of data from all influenza viruses and the coronavirus causing COVID-19. This includes genetic sequence 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.*
>
> *The Initiative ensures that open access to data in GISAID is provided free-of-charge to all individuals that agreed to identify themselves and agreed to uphold the GISAID sharing mechanism governed through its Database Access Agreement" (from GISAID web site).*

Many governments or regulatory or supranational health institutions use GISAID data as is or filtered as they need. One of them is Cov-Lineage (https://cov-lineages.org/) which

Published: March 25, 2022

collects and describes all variant characteristics according to Pango nomenclature. All sequences are downloaded daily from GISAID and processed to eliminate sequences with uncertain or not specified collection data, incomplete sequences, and those that present more than 5000 nucleotide substitutions or more than 500 base pair contiguous deletions. Moreover, sequence rates between locations are highly variable and the S-gene target often fails during PCR amplification. So, sequencing techniques can heavily bias the samplings (https://outbreak.info/situation-reports/methods).

As a group of researchers engaged in the characterization of Spike glycoprotein mutations and in the development of models capable of predicting future possible stable mutations, we have encountered several complications during the analysis of the genomic sequences.

Although GISAID provides clear indications necessary for the submission of sequences and related information, without a persuasive control, a large amount of incoming data can compromise the data quality, making it largely unfit for use and a data cleaning procedure necessary, as in our case.

We have considered 6 908 513 sequences of Spike glycoprotein updated on January 12, 2021, downloaded in Fasta format. Each sequence is characterized by a header that contains some useful information like the sequence ID, date of collection, sequencing laboratory, and world region, although others could be added like the Pango ID or WHO label that indicates the lineage to which the sequence is assigned. The current sequence updates on GISAID rely on the single laboratories involved in the sequencing, and it is operator dependent. There are some concerns about this procedure: for example, the virus name and its passage details are inserted manually, and this could introduce some problems with automated analysis, due to spelling errors. In our analysis, we found, for example, 44 653 sequences that did not have "original" (the analysis is case insensitive) passage details, but 29 860 of them were annotated as "orginal".

It is worth mentioning that the uploaded sequences can have different lengths, ranging from a maximum of 59 503 nucleotides to a minimum of a few dozen. The sequences have an approximate length of 29k nucleotides, and the wildtype sequence length is 29 891. This leads to the presence of low-quality sequences defined by GISAID itself as "sequences represented by NNNNs in the nucleotide genome and codons with ambiguous bases and translated into X". Overall, countries in different geographical regions have submitted a different number of sequences with an average low-quality sequences rate of 40%, as shown in Figure 1. The letter X, in the case of the amino acid sequences, or the letter N, in the case of the nucleotide ones, that indicates a low-quality region, after a more deepen analysis, may be due to deletions. The deletions are noted by simply shortening the sequence length.

This behavior could lead to difficulties in future analysis of comparison and mutation detection and could be handled by fixing a minimum of length for the sequence to upload or by using placeholder for deletions and fixed wildtype length. Doing so could help with a faster alignment of the sequence in the fasta format and easier detection of deletions and insertions (if the sequence is longer than the wildtype, then there is at least an insertion).

Since it is not possible to have a custom download of the entire database that allows one to filter the data, excluding, for example, sequences with low quality or without a correct collection data or not human sequences, we have developed a data cleansing procedure, drastically reducing the number of useful sequences, as shown in Figure 2.

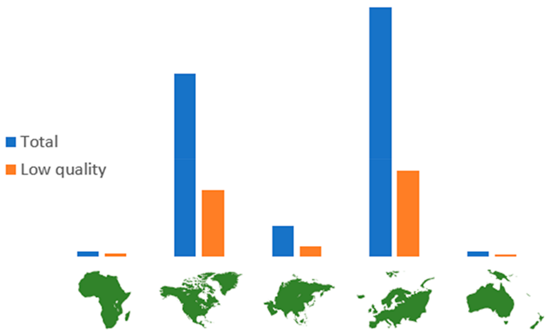


Figure 1. Area distribution of low-quality regions: blue represents the total number of submitted sequences, while the orange represents low quality sequences. Most of the sequences are derived from America, Europe, and Asia, while Africa and Oceania are the regions with the highest rate of low-quality sequences.

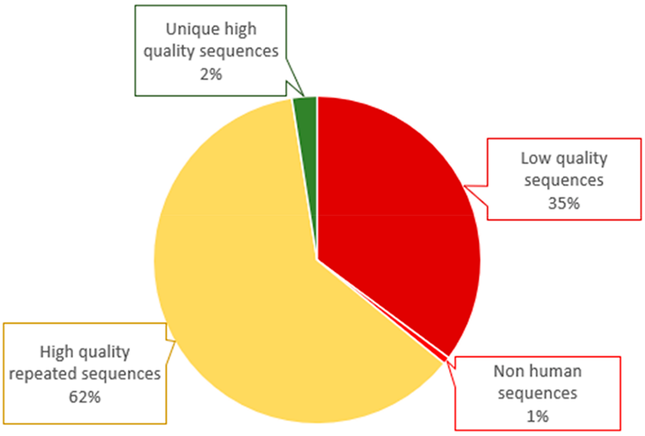


Figure 2. Data cleansing procedure: Low-quality sequences (sequences that present at least one X) and nonhuman sequences are excluded. High quality repeated sequences represent 62% of the distribution, and unique high-quality sequences, used for our research purposes, represent only 2%.

How can we solve these problems by providing quality data that can be used directly by everyone? Blockchain technology could be a fast, inexpensive, and effective solution.

The diffusion at global level of the blockchain technology (+39% increase in the number of projects between 2020 and 2021) is driven by the necessity of reaching certain standards of efficiency in areas like payments, the transparency of the supply chain, and the authenticity and recognizability of any kind of information in a safe and quick way. In recent years, blockchain technology has been applied in different sectors such as food[1,2] and health care. It has recently been evaluated as a useful tool in the management of the COVID-19 emergency[3] that can be used, for example, for contact tracing or the administration of vaccines.[4,5]

Blockchain technology can substitute every process of authentication of information (which would otherwise be slow and not completely reliable), given that it guarantees the origin and immutability of the information in the exact same moment when it is adopted. The instant confirmations of origin and authenticity can become a fundamental capability for the “guaranteed and transparent” use of information, like in our case of “Spike sequences”, eliminating the time dedicated to their validation before their employment and drastically reducing the time for their elaboration.

In particular, the GISAID database will not exist anymore in a physical and known place where it is controlled by the operators, but it will exist in the “logical place” managed by the blockchain, therefore becoming public and managed by third parties. In addition, the information is distributed in different servers, and therefore, its physical preservation is guaranteed even if a machine or part of it defaults (guaranteed backup). Information is encrypted through the calculation of a hash, and it is not modifiable because this technology does not allow any kind of modifications. The only possible modifications consist of providing a new piece of information with an explicit indication of which one becomes obsolete. The depositor is recognized, and their identity is indistinguishable from the information they provide because they are recognized with certainty at the time of deposit. The information is qualified because it is public and accessible to whoever has the right. For this reason, it is called a “public ledger”.

From a practical point of view, every single lab will produce a certain piece of information, which will be deposited using specific software interfaces able to interact with any software package used for local management, in a specific blockchain that will be managed by third parties. The information, deposited by those authorized to do so is encrypted through the calculation of the hash and written and inserted within the nodes (at last three). At the end of this process, the information will be unmodifiable and available (only for consultation) to every person who has the right to it.

It is useful to highlight a note regarding the process of exchanging and depositing information. It is reasonable to predict that every lab will be provided with software licenses to manage its activities with functional characteristics based on its activity and on blockchain technology. Indeed, it is crucial to create not only an efficient blockchain but also instruments that allow the efficient and safe exchange of data.

Different software solutions have recently appeared on the market, and they offer interesting characteristics that use, for instance, virtual databases able to publish logic views managed through physical tables, with advantages in terms of speed and safety of transactions. These solutions have been developed to avoid any element of fragility on the entire process through which information passes, from the laboratories to its deposition in the blockchain.

Today, international competitors like IBM, Microsoft, and AWS (but also others) are able to offer reliable infrastructures at moderate prices, making the creation of databases that use this technology not too expensive. Furthermore, it is possible to use graphical interfaces, suitably developed, which make the use of this technology accessible to all.

As far as the limit of the blockchain is concerned, many studies have been carried out, especially during the last 2 years, to reduce the time of execution of validation. These studies have led to the experimentation and diffusion of two technologies like the “BLS digital signature”, also known as Boneh–Lynn–Shacham, and “sharding”, that drastically reduce the space to store the nodes. Other solutions have also been found, and the result is impressive if compared to options available a couple of years ago when the latency time was measurable in minutes but now can be measured in less than 1 s.

Last but not least is the possibility of realizing private blockchains where the owner can establish its rules and its dimension without losing the characteristics of authenticity and immutability guaranteed by third parties. Furthermore, the increase in the use of this technology in areas like the production of art pieces (through NFTs or Nun-Fungible Tokens), in the video game industry, or in the acquisition of images of sports events (always through NFTs) guarantees the constant improvement in performance with a constant decrease in costs.

All these elements show that today a generic database protected by the blockchain is possible and the only obstacle lies in the willingness and/or capacity to build it.

AUTHOR INFORMATION

Corresponding Author

Pietro Cozzini – *Molecular Modelling Lab, Food & Drug Department, University of Parma, 43124 Parma, Italy;* orcid.org/0000-0002-4826-8108; Email: pietro.cozzini@unipr.it

Authors

Federica Agosta – *Molecular Modelling Lab, Food & Drug Department, University of Parma, 43124 Parma, Italy*

Greta Dolcetti – *Greta Dolcetti, Department of Mathematical, Physical and Computer Sciences, University of Parma, 43124 Parma, Italy*

Gianfranco Righi – *Gianfranco Righi, UNITEAM srl, 40138 Bologna, Italy*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsmedchemlett.2c00077

Author Contributions

This viewpoint manuscript has origin from research work on COVID-19 mutation prediction; thus, all the authors contributed equally to the work.

Notes

Views expressed in this viewpoint are those of the author and not necessarily the views of the ACS.
The authors declare no competing financial interest.

REFERENCES

(1) Galvez, J. F.; Mejuto, J. C.; Simal-Gandara, J. Future challenges on the use of blockchain for food traceability analysis. *TrAC, Trends Anal. Chem.* **2018**, *107*, 222–232.

(2) Motta, G. A.; Tekinerdogan, B.; Athanasiadis, I. N. Blockchain Applications in the Agri-Food Domain: The First Wave. *Front. Blockchain* **2020**, 3, 6.

(3) Alsaed, Z. Role of blockchain technology in combating COVID-19 crisis. *Appl. Sci. (Switz.)* **2021**, *11*, 12063.

(4) Ng, W. Y. Blockchain applications in health care for COVID-19 and beyond: a systematic review. *The Lancet Digital Health* **2021**, *3* (12), e819–e829.

(5) Shah, H.; Shah, M.; Tanwar, S.; Kumar, N. Blockchain for COVID-19: a comprehensive review. *Personal Ubiquitous Comput.* **2021**, DOI: 10.1007/s00779-021-01610-8.