SENSIBLE: implementing data-driven early warning systems for future viral epidemics

Anna Bernasconi[1,*], Matteo Chiara[2], Tommaso Alfonsi[1] and Stefano Ceri[1]

[1]*Department of Electronics, Information and Bioengineering – Politecnico di Milano, Milan, Italy*
[2]*Department of Biosciences – Università degli Studi di Milano, Milan, Italy*

Abstract

In March 2020, COVID-19 was declared a global pandemic, spurring intense research efforts. Genomic surveillance emerged as a crucial defense against the virus. Variants of concern arise from epidemiologically relevant mutations, posing challenges in disease control. Global pandemics are exacerbated by habitat loss, urbanization, and globalization, facilitating disease spread. The SENSIBLE project, supported by the Italian Ministry of University and Research through NextGeneration EU funding, aims to develop methods for analyzing viral genomes and implement early warning information systems based on data-driven analysis, exploiting data from past epidemics for validation. SENSIBLE will create an integrated framework for genomic surveillance, utilizing data-driven and knowledge-based analyses. By leveraging knowledge from COVID-19, it seeks to enhance understanding of viral pathogens and aid healthcare decision-making.

Keywords

Bioinformatics, Big Data, Data Science and Analytics, Biomedical Ontologies, Genomic Surveillance, Pathogens Evolution, IS in the post-COVID world

Project Information Project Name: Small-data Early warNing System for viral pathogens In puBLic hEalth (SENSIBLE); Funding Agency: Italian Ministry of University and Research (MUR) funded by NextGeneration EU; URL: https://sensible-prin.github.io/; Running Period: 01.12.2023 – 30.11.2025; Principal Investigator: Anna Bernasconi (Politecnico di Milano); Co-Principal Investigator: Matteo Chiara (Università di Milano).

1. Project overview

In March 2020, COVID-19 was declared a global pandemic. The research community mounted an unprecedented effort to understand the disease and its etiological agent, deliver effective diagnostics, plan vaccination programs, and inform decision-making and public health policies.

The "recent" developments in high throughput DNA sequencing technologies [1] and the associated reduction in costs strongly favored the availability of genomic sequences, derived from different specimens, and collected at different locales. This, in turn, enabled the development

Research Projects Exhibition @CAiSE'24, June 3-7, 2024, Limassol, Cyprus

[*] Corresponding author.

anna.bernasconi@polimi.it (A. Bernasconi); matteo.chiara@unimi.it (M. Chiara); tommaso.alfonsi@polimi.it (T. Alfonsi); stefano.ceri@polimi.it (S. Ceri)
0000-0001-8016-5750 (A. Bernasconi); 0000-0003-3983-4961 (M. Chiara); 0000-0002-4707-8425 (T. Alfonsi); 0000-0003-0671-2415 (S. Ceri)

of “genomic surveillance” systems [2], that allow us to study the spread and evolution of pathogens through the comparison of genomic sequences. Genomic surveillance was universally recognized as a first line of defense to contrast the pandemic. All viruses mutate as they replicate and spread in a population; the majority of mutations are not relevant from an epidemiological perspective. However, epidemiologically relevant mutations might confer a selective advantage and are rapidly fixed in the pathogen genome, leading to the emergence of “variants of interest” or “variants of concern”. Global pandemics are an accelerating threat: loss of habitats, urbanization, and globalization create an environment conducive to infectious disease outbreaks and spread. In the wake of climate change, disease vectors such as Asian tiger mosquitos are now endemic in Europe – and linked to outbreaks of diseases (e.g., Zika virus, Chikungunya virus).

At the end of 2022, we proposed SENSIBLE (Small-data Early warNing System for viral pathogens In puBLic hEalth), in the context of the Italian PRIN PNRR 2022 funding, targeting the ERC fields PE6 “Web and information systems, data management systems, information retrieval and digital libraries, data fusion” along with “Bioinformatics, bio-inspired computing, and natural computing”.

The SENSIBLE project was selected for funding, lasting two years, starting in December 2023. SENSIBLE aims to leverage the knowledge gained on COVID-19 to build *novel methods to handle and analyze pathogens’ genome sequencing data in current and future viral epidemics*, and *implement an early warning system based on data-driven analysis*. The project’s consortium is composed of two partners with complementary competencies in data management and tool development for life sciences-related domains (Politecnico di Milano) and genomics, bioinformatics, and viral evolution (University of Milan).

During the project frame, SENSIBLE will develop an integrated framework for genomics surveillance of human pathogens, based on the integration of 1) data-based analysis to summarize patterns of evolution through space and time; 2) data and knowledge-based analysis (retrieval, computation, or prediction) to formulate testable biological hypotheses and identify epidemiologically relevant evolutionary events (positive selection, change in protein function/affinity, immune escape). The framework will be developed and validated using a selection of use cases from COVID-19; a final assessment will be performed on independent data from the recent Monkeypox (2022), Zika (2015-2016), and Ebola (2013-2016) epidemics.

SENSIBLE will advance the state of the art in understanding the various facets of genomic surveillance, depending on the available data and the domain context. We will merge the experience gathered on COVID-19 with the considerable knowledge corpus that has become openly available for viral pathogens research. The project results will have a substantial impact on the early characterization of novel viral pathogens and their dangerousness in terms of prevalence, infectivity, and transmissibility. More importantly, the framework developed by SENSIBLE will provide a highly useful tool to assist decision-makers in healthcare.

2. Project positioning and objectives

Over 16 million genomic sequences of the etiological agent-SARS-CoV-2- have been determined in a span of less than 4 years. This large body of data enabled the study and monitoring of viral evolution at an unprecedented scale and allowed scientists to better understand SARS-CoV-2 variants and the risks that they pose [3, 4]. Considerable efforts have been dedicated to building surveillance systems, to assist in the identification of viral variants and their potential impact and to assist decision-making in healthcare [5].

Several tools that leverage large collections of SARS-CoV-2 genome sequences, as available from public repositories (GISAID [6] and NCBI Virus GenBank [7]), have been introduced in the last two years (see CoV-Spectrum [8], COVID-19 CG [9], Outbreak.info [10], ViruClust [11] and VariantHunter [12], among others). Attempts to build completely automated early warning systems, based on the availability of "early" sequences as they became available through public repositories, have been proposed as well. Such methods are mainly based on unsupervised machine learning algorithms and exploit a range of features to describe the epidemiological trends of the epidemic. However, these tools found limited application in the context of the COVID-19 pandemic, where strategies for genomic surveillance were, in the majority of cases, based on the retrospective analysis and observation of genomic data.

Unfortunately, a monitoring strategy that is reliant only on big data is coarse-grained and limited in its potential. Such a strategy presents three main limitations: first, the need for large amounts of data to produce statistically relevant evidence; second, the need to access the data within a short time-frame from its collection, which is not always possible; third, the lack of integration of epidemiological data with known characteristics of the virus, which is often disregarded.

New, emerging viral pathogens – that may arise unexpectedly – will not provide the ideal settings for "big-data"-based genomic surveillance. SENSIBLE aims to fill this gap by eliminating the dependence on big and timely collected data and by integrating a wide range of annotations and features that can be gathered from public databases or computed based on similar scenarios. We propose a systematic approach based on the presence of "small" datasets and conceptually distinct modules, each gathering one layer of information, that – when merged – can deliver a broad understanding of complex mechanisms of viral evolution.

Specifically, SENSIBLE aims to empower pandemic preparedness by building a general information system for the genomic surveillance of pathogens in current and future viral human epidemics. We will address the following three objectives:

1. Derive effective methods for data-driven identification of emerging viral pathogens;
2. Build an objective framework for genomic surveillance in current and future epidemics;
3. Implement an early warning system, to assist decision-making in healthcare.

The latter objective will lead to the main outcome of the project, acting as early as possible, to identify emerging viral pathogens and/or novel lineages of known pathogens that might pose an immediate risk to human health.

3. Methodologies

SENSIBLE will explore and harness data from different domains of interest, including: analyses of available data, mapping of equivalent/matched information from similar pathogens, computation or prediction of novel features and properties of the virus under study. The framework developed by SENSIBLE will feature four main – conceptually distinct – tasks:

1. **Data-driven analysis for the study of pathogens' evolution.** Both project partners have relevant expertise in the development of methods and tools for the genomic surveillance of pathogens based on large scale/big data [13, 14, 15]. Techniques based on sub-sampling and bootstrapping can be applied to extend the range of applications of these methods, identify minimal subsets of actionable data, and evaluate their validity and robustness.

2. **Data and knowledge-based analysis.** To translate viral evolutionary dynamics into a collection of "epidemiologically-relevant" annotations of the viral genome, different sources of information will be gathered:

   - functional annotations of genomic elements;
   - highly conserved or hyper-variable genomic regions;
   - sites under positive/negative selection;
   - data on predicted and/or validated T and B-cell epitopes (if/when available);
   - alteration in protein functions (e.g., binding) based on interaction with hosts' proteins.

   Annotations available from existing resources will be retrieved and integrated into an internal knowledge base; missing data will be computed (via bioinformatics tools) or predicted (via automatic learning procedures).

3. **Ranking and prioritization.** Based on the evolutionary observations derived in (1) and the detailed functional annotations (retrieved, computed, or predicted) from (2), a prioritization score will be computed to assign a "level of concern" to emerging viral pathogens and or to novel viral lineages. The score will be exploited to develop a ranking system, which will be evaluated according to the heuristics to be developed.

4. **Validation and testing.** The initial development and setup of SENSIBLE will be performed on a selection of use cases from the COVID-19 pandemic. Monkeypox (2022), Zika (2015-2016), and Ebola (2013-2016) will be used to showcase the system and provide an unbiased evaluation.

4. Expected results

SENSIBLE aims to shift current paradigms in pathogens' genomic surveillance. We move away from the traditional approach based purely on monitoring the prevalence of viral variants and lineages, based on big data and, instead, we propose an integrated approach, that by leveraging a collection of different key evolutionary and epidemiological features will allow a more complete understanding of an epidemic, producing more informative results, and without the need for large amounts of data.

We will tackle two key epidemiological questions and identify key metrics for raising alerts and early warnings in both scenarios:

- *Minimal actionable data.* What is the minimal amount of data production/availability required to set up an effective surveillance system? Can genomic surveillance be applied even in scarce/low-resource settings? Notwithstanding the recent experience with COVID-19, these questions remain largely unanswered. **We aim to provide resource-aware recommendations/guidelines and quantifiable metrics to assist health authorities in the set-up of "minimal" pathogen surveillance systems.**

- *Prioritization/ranking of emerging pathogens.* If a new mutation or pattern of mutations arises in a human pathogen, how does this impact its epidemiological features? Our first focus in this case is changes that may provoke increased transmissibility, contagiousness, infectivity, or immune system evasion. **We plan to build a scoring system to rank and prioritize emerging virus/viral variants with enhanced epidemiological features.**

5. Relevance to CAiSE

The early warning system that SENSIBLE will produce will be the core of a future genomic surveillance information system to be employed at the national and international levels for monitoring any kind of evolutionary phenomenon that is based on the collection of several data points to be integrated with domain-specific knowledge-based modules. Our project shares several topics of the CAiSE conference such as: "*Big Data, Data Science and Analytics*", a fundamental discipline to understand the mechanisms of evolution in life sciences domains; "*Artificial Intelligence and Machine Learning*", necessary to the development of the bioinformatics data-driven analysis framework to identify interesting mutations, variants, evolutionary patterns; "*Ontologies and Ontology Engineering*", partially involved as our knowledge-based analysis will rely on domain-relevant biomedical ontologies and terminologies and on the achievement of interoperability among specialized resources that are currently lacking any connection (see past work in this area [16, 17]; and "*IS in the post-COVID world*" / "*IS for healthcare*", both of which demonstrate relevant applications of Information Systems engineering in the context of life sciences.

The results of SENSIBLE will strongly impact future-generation healthcare systems which will gain an improved awareness of the evolutionary mechanisms of surrounding viral pathogens.

6. Project status

In the first 4 months (of the 2-year span of the project), efforts focused on sourcing datasets and reference databases while delineating essential framework functionalities, with emphasis especially on SARS-CoV-2 recent evolution (i.e., recombinant lineages), monkeypox, and influenza viruses. Attention was devoted to gathering datasets and annotations from public repositories.

The team initiated exploration into specific instances such as recombined SARS-CoV-2 lineages (with a publication on Nature Communications [18]), the accumulation of mutation in specific regions called epitopes [19], and expanded the techniques so far crafted to the new domain of influenza viruses, with a focus on avian flu. Its pandemic potential would likely escalate [20] if it were to become transmissible from mammals to humans and –eventually– from humans to humans. Currently, we are mapping software methods to functionalities like variation, recombination, and reassortment identification, all important evolutionary mechanisms for flu viruses.

7. The potential impact of SENSIBLE

"*Messieurs, c'est les microbes qui auront le dernier mot.*"

– Louis Pasteur, 19th-century microbiologist

Viruses are among the most important causes of morbidity and mortality worldwide. Rapid diagnosis and implementation of effective therapies/strategies to limit their spread represents a major challenge for Public Health systems. Habitat degradation worldwide, compounded by global connectivity including the increasing levels of contact between humans and animal reservoirs for zoonoses, means that the readiness to tackle future infectious pathogens and the ability to maintain public health and functional economies will represent an important challenge for the future wellbeing of society.

Additionally, the COVID-19 pandemic has highlighted the epidemic/pandemic potential of emerging zoonotic viruses, due to naive immunity in the general population. Several new viral infectious diseases with epidemic potential – that could threaten global health and security – have emerged over the past years. The cost of not preparing for epidemics was estimated at $60 billion in 2016 [21], an estimate dwarfed by the projected cost of the current coronavirus pandemic, which is by one estimate over $16 trillion in the US alone [22].

Methods for the genomic surveillance of emerging pathogens have been of fundamental importance to contrast the spread of COVID-19 and have provided a first line of defense against the pandemic. However useful, currently available systems suffer from inherent limitations, among which the most important ones are: requirement for the availability of big/large scale data need for the usage of multiple distinct tools and lack of interoperability limited predictive power.

One of the most important consequences is that – at the time being – novel and possibly more dangerous variants of a known viral pathogen can only be identified in hindsight and only from countries/regions in the world where a sufficient amount of data is available.

The vision of SENSIBLE is to overcome such limitations, by promoting a shift in paradigm in the methods and concepts applied to the genomic surveillance of pathogens. We move away from the traditional approach based purely on the observation of big data and their trends; instead, we propose a more proactive approach, where –by integrating and leveraging key evolutionary and epidemiological features– enable the prediction of key epidemiological events, and raise early warnings, without the need for large amounts of data.

SENSIBLE will have significant **technological impacts**. From a data management and integration perspective, we expect to advance the knowledge in the interoperability of virus-related data, composing a database that will support both sequences and all the knowledge that is connected to it. From a **data science** perspective – in conjunction with bioinformatics techniques – we expect the introduction of new methods that will inform on viral transmissibility, infectivity and prevalence. The project will also advance the state of the art in the genomic surveillance field by 1) enabling monitoring also in "small data" scenarios; 2) promoting a more proactive strategy based on early warnings, rather than data observation; 3) facilitating the interoperability of data and methods for the surveillance of pathogens. Moreover, SENSIBLE will be able to make a valuable contribution to the **decision making** perspective: we will provide useful support to decision makers, fully open and adaptable to specific needs; in particular we expect an impact in public health management and healthcare.

The usability of the methods and services developed by SENSIBLE will be demonstrated through a set of use cases designed to reflect the continuously evolving research questions in a pandemic. All the methods will be tested and validated by applying our framework to different use-cases, based on real world data. This approach will showcase potential applications of our framework and the feasibility of our approach to stakeholders and prospective users.

We envisage that SENSIBLE might provide a highly useful tool for stakeholders in National Health Systems and/or Regional Health authorities, including for example the Italian National Institute of Health (ISS, Istituto Superiore di Sanità) and/or equivalent authorities in other countries.

In conclusion, the SENSIBLE framework will provide an innovative, inter/multi-disciplinary ecosystem for the genomic surveillance of pathogens, which will contribute to enhancing our preparedness (encompassing decision- and policy-making, and behaviors) for infectious disease epidemics. Beyond providing innovative tools and methods, SENSIBLE will provide researchers and stakeholders with technical solutions and skills to address the very concrete threats posed by emerging viral pathogens, and/or novel variants of known pathogens leading to improved understanding of the occurrence and spread of these, including their effect on disease severity and vaccine effectiveness. Altogether, our project will contribute to building global pandemic preparedness so that ‘microbes do not have the last word’.

Acknowledgements.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

The work is supported by Ministero dell’Università e della Ricerca (PRIN PNRR 2022 “SENSIBLE” project, n. P2022CNN2J), funded by the European Union, Next Generation EU, within PNRR M4.C2.1.1. Politecnico di Milano CUP D53D23017400001; Università degli Studi di Milano CUP G53D23006690001. PI Anna Bernasconi, co-PI Matteo Chiara.

References

[1] S. C. Schuster, Next-generation sequencing transforms today’s biology, Nature methods 5 (2007) 16.

[2] J. L. Gardy, N. J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system, Nature Reviews Genetics 19 (2018) 9–20.

[3] S. W. Lo, D. Jamrozy, Genomics and epidemiological surveillance, Nature Reviews Microbiology 18 (2020) 478–478.

[4] L. Subissi, A. von Gottberg, L. Thukral, N. Worp, B. B. Oude Munnink, S. Rathore, L. J. Abu-Raddad, X. Aguilera, E. Alm, B. N. Archer, et al., An early warning system for emerging SARS-CoV-2 variants, Nature Medicine 28 (2022) 1110–1115.

[5] J. A. Plante, B. M. Mitchell, K. S. Plante, K. Debbink, S. C. Weaver, V. D. Menachery, The variant gambit: COVID-19’s next move, Cell host & microbe 29 (2021) 508–515.

[6] Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data–from vision to reality, Eurosurveillance 22 (2017) 30494.

[7] E. W. Sayers, M. Cavanaugh, K. Clark, K. D. Pruitt, S. T. Sherry, L. Yankie, I. Karsch-Mizrachi, GenBank 2024 update, Nucleic Acids Research 52 (2024) D134–D137.

[8] C. Chen, S. Nadeau, M. Yared, P. Voinov, N. Xie, C. Roemer, T. Stadler, CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants, Bioinformatics 38 (2022) 1735–1737.

[9] A. T. Chen, K. Altschuler, S. H. Zhan, Y. A. Chan, B. E. Deverman, COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest, Elife 10 (2021) e63409.

[10] K. Gangavarapu, A. Latif, J. Mullen, et al., Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations, Nature Methods 20 (2023) 512–522.

[11] L. Cilibrasi, P. Pinoli, A. Bernasconi, A. Canakoglu, M. Chiara, S. Ceri, ViruClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time, Bioinformatics 38 (2022) 1988–1994.

[12] P. Pinoli, A. Canakoglu, S. Ceri, M. Chiara, E. Ferrandi, L. Minotti, A. Bernasconi, VariantHunter: a method and tool for fast detection of emerging SARS-CoV-2 variants, Database 2023 (2023) baad044.

[13] A. Bernasconi, L. Mari, R. Casagrandi, S. Ceri, Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence, Scientific Reports 11 (2021) 21068.

[14] M. Chiara, D. S. Horner, C. Gissi, G. Pesole, Comparative genomics reveals early emergence and biased spatiotemporal distribution of SARS-CoV-2, Molecular biology and evolution 38 (2021) 2547–2565.

[15] M. Chiara, D. S. Horner, E. Ferrandi, C. Gissi, G. Pesole, Unsupervised classification of SARS-CoV-2 genomic sequences uncovers hidden genetic diversity and suggests an efficient strategy for genomic surveillance, bioRxiv (2021) 2021–06.

[16] G. Guizzardi, A. Bernasconi, O. Pastor, V. C. Storey, Ontological unpacking as explanation: the case of the viral conceptual model, in: Conceptual Modeling: 40th International Conference, ER 2021, Virtual Event, October 18–21, 2021, Proceedings 40, Springer, 2021, pp. 356–366.

[17] A. Bernasconi, G. Guizzardi, O. Pastor, V. C. Storey, Semantic interoperability: ontological unpacking of a viral conceptual model, BMC bioinformatics 23 (2022) 491.

[18] T. Alfonsi, A. Bernasconi, M. Chiara, S. Ceri, Data-driven recombination detection in viral genomes, Nature Communications 15 (2024) 3313.

[19] R. Al Khalaf, A. Bernasconi, P. Pinoli, Supporting data for "Systematic analysis of SARS-CoV-2 Omicron subvariants' impact on B and T cell epitopes", 2024. URL: https://doi.org/10.5281/zenodo.10517709.

[20] D. M. Morens, J. Park, J. K. Taubenberger, Many potential pathways to future pandemic influenza, Science Translational Medicine 15 (2023) eadj2379.

[21] P. Sands, C. Mundaca-Shah, V. J. Dzau, The neglected dimension of global security – a framework for countering infectious-disease crises, New England Journal of Medicine 374 (2016) 1281–1287.

[22] D. M. Cutler, L. H. Summers, The COVID-19 pandemic and the $16 trillion virus, Jama 324 (2020) 1495–1496.