Amino Acids (2024) 56:3
https://doi.org/10.1007/s00726-023-03374-2

ORIGINAL ARTICLE

Structure determination needs to go viral

Matheus de Bastos Balbe e Gutierres[1] · Conrado Pedebos[1] · Paula Bacaicoa-Caruso[1] · Rodrigo Ligabue-Braun[1,2]

Received: 29 May 2023 / Accepted: 14 November 2023

Abstract
Viral diseases are expected to cause new epidemics in the future, therefore, it is essential to assess how viral diversity is represented in terms of deposited protein structures. Here, data were collected from the Protein Data Bank to screen the available structures of viruses of interest to WHO. Excluding SARS-CoV-2 and HIV-1, less than 50 structures were found per year, indicating a lack of diversity. Efforts to determine viral structures are needed to increase preparedness for future public health challenges.

Keywords Virus · PDB · X-ray · NMR · Cryo-EM · Pandemic

Introduction

The emergence of SARS-CoV-2 as the causative agent of the COVID-19 pandemic brought focus back to the emergence and re-emergence of infectious diseases as global health threats, especially those caused by viruses (Ciotti et al. 2020; da Silva et al. 2022). There is no established number of human-infecting virus species (and even the term 'species' might be inadequate), but a current estimate is of at least 219 virus species, under 23 families (Woolhouse et al. 2012). Despite these figures being estimates, data modeling and extrapolation propose from 513 novel viruses to 827,000 viruses as potentially human-infecting, from a universe of 1.67 million unknown viruses (Chatterjee et al. 2021). At least two thirds of these are of zoonotic origin, reinforcing the chance of transmission from farm or wild animals to humans (Pandit et al. 2022). In addition to that, half the viruses that can infect humans are also transmissible among humans, and half of those are able to generate more than one secondary case after infection (i.e. $R_0 > 1$) (Woolhouse et al. 2012).

The potential risk of emerging infections led the World Health Organization (WHO) to prioritize some diseases for which there is epidemic potential and/or there are no or insufficient countermeasures (WHO 2023). The WHO priority list encompasses COVID-19, Crimean-Congo hemorrhagic fever, Ebola virus disease and Marburg virus disease, Lassa fever, Middle East respiratory syndrome coronavirus (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS), Nipah and henipaviral diseases, Rift Valley fever, Zika, and "Disease X". WHO defines Disease X as "the knowledge that a serious international epidemic could be caused by a pathogen currently unknown to cause human disease", for which preparedness for other diseases could also be relevant (Chatterjee et al. 2021; WHO 2023).

The availability of protein structures is one of the basic requirements for rational drug design (Hol 1986; Sliwoski et al. 2013), especially the structure-based drug design route (Batool et al. 2019). In this strategy, the molecular target (generally a protein) is inspected in terms of its structure, providing stereoelectronic insights for the development of drug candidates (Mandal et al. 2009; Mavromoustakos et al. 2011). To obtain a protein structure experimentally (a process known as structure determination), an analytical technique is employed to solve three-dimensional atomic

Handling editor: A. de Brevern.

Matheus de Bastos Balbe e Gutierres and Conrado Pedebos have contributed equally to this work.

✉ Rodrigo Ligabue-Braun
rodrigolb@ufcspa.edu.br

[1] Programa de Pós-Graduação em Biociências (PPGBio), Universidade Federal de Ciências da Saúde de Porto Alegre - UFCSPA, Porto Alegre, Rio Grande do Sul, Brazil

[2] Departamento de Farmacociências, Universidade Federal de Ciências da Saúde de Porto Alegre - UFCSPA, Porto Alegre, Rio Grande do Sul, Brazil

Published online: 29 January 2024

coordinates. These commonly include X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (Cryo-EM) (Stollar and Smith 2020).

In the case of SARS-CoV-2, there was an unprecedented output of resolved viral protein structures in a very short period of time (Lynch et al. 2021). A similar effort has only been seen for HIV, albeit on a much smaller scale in a much longer period of time (Engelman and Cherepanov 2012). Thus, considering the viral abundance (and its associated risks) in one hand, and the need for protein structure determination in the drug development pipelines in the other, the aim of this study was to assess how well represented is the viral diversity in terms of its protein structures as deposited in the RCSB Protein Data Bank. We were able to confirm a clear skew towards SARS-CoV-2 protein structures, followed by HIV, but at a very different speed of deposition. Other priority-level viruses are underrepresented, reinforcing the current need for structural determination focused on potentially (re) emergent viral agents.

Methods

On September 10th, 2022, the RCSB Protein Data Bank (rcsb.org) (Berman et al. 2000; Burley et al. 2021) was queried for viral protein structures, using the "Browse by Annotations" option under the Search tab. By limiting results by Source Organism as "Viruses", 10,239 entries were located and a tabular report was generated via databank interface. Custom Python scripts (Supplementary File 1) were developed for extracting information under analysis in this work (number of deposited files per virus, structural determination method employed, average resolution, date of deposition) and for graphically representing the obtained results (as amount of deposited structures by year).

Results and discussion

The original tabular report obtained from RCSB PDB is shown in Supplementary File 2. It corresponds to 0.18% of the entire database (5,717,483 total). Since we decided to focus on the WHO priority list of human viruses, the original set containing all 10,239 virus entries was filtered, leading to 5662 entries. The taxonomy definitions for each virus are listed on Supplementary File 3. As can be seen on Table 1, the results are dominated by SARS-CoV-2 and HIV 1 depositions. It is possible to observe that Cryo-EM became the second most prevalent structural determination technique, especially in the case of SARS-CoV-2 where it far exceeds NMR in number of structures elucidated. Note that 23 deposited entries were obtained by other methods than X-ray diffraction, nuclear magnetic resonance spectroscopy, or Cryo-EM, and are not represented in this list.

In order to inspect potential trends in viral proteins structural determination, we plotted the number of structures deposited per virus in the databank by year (Fig. 1A). As expected, the number of structures from the SARS-CoV-2 virus has increased drastically in the last three years, due to the efforts in studying the virus during the COVID-19 pandemic. Values reached more than a thousand structures in just about 1 year (2020) and were kept up above 500 structures in subsequent years. When we remove SARS-CoV-2 structure entries from the plot (Fig. 1B), the number

Table 1 Structural entries for the WHO priority list of viruses

Virus	X-Ray	Cryo-EM	NMR	Avg XR res.	Avg EM res.	Total
SARS-CoV-2	1688	916	13	1.86	3.42	2626
HIV 1	1811	242	115	2.23	4.82	2182
SARS Virus	189	32	28	2.12	3.72	249
Zika virus	141	22	2	2.11	7.41	165
MERS virus	123	22	0	2.30	3.77	145
Ebola virus	67	22	1	2.39	4.24	90
Lassa fever virus	30	18	1	2.31	4.09	49
HIV 2	26	0	8	2.15	–	34
Rift Valley fever virus	22	6	1	2.51	10.79	29
Nipah virus	20	6	0	2.44	3.48	26
Henipa virus	25	1	0	2.63	2.80	26
Crimean-Congo hemorrhagic fever virus	20	1	1	2.27	2.80	22
Marburg virus	18	1	0	2.48	3.10	19
						5662

X-Ray X-ray diffraction, *Cryo-EM* cryogenic electron microscopy, *NMR* nuclear magnetic resonance spectroscopy, *Avg XR res.* average x-ray resolution (Å), *Avg EM res.* average cryo-EM resolution (Å)

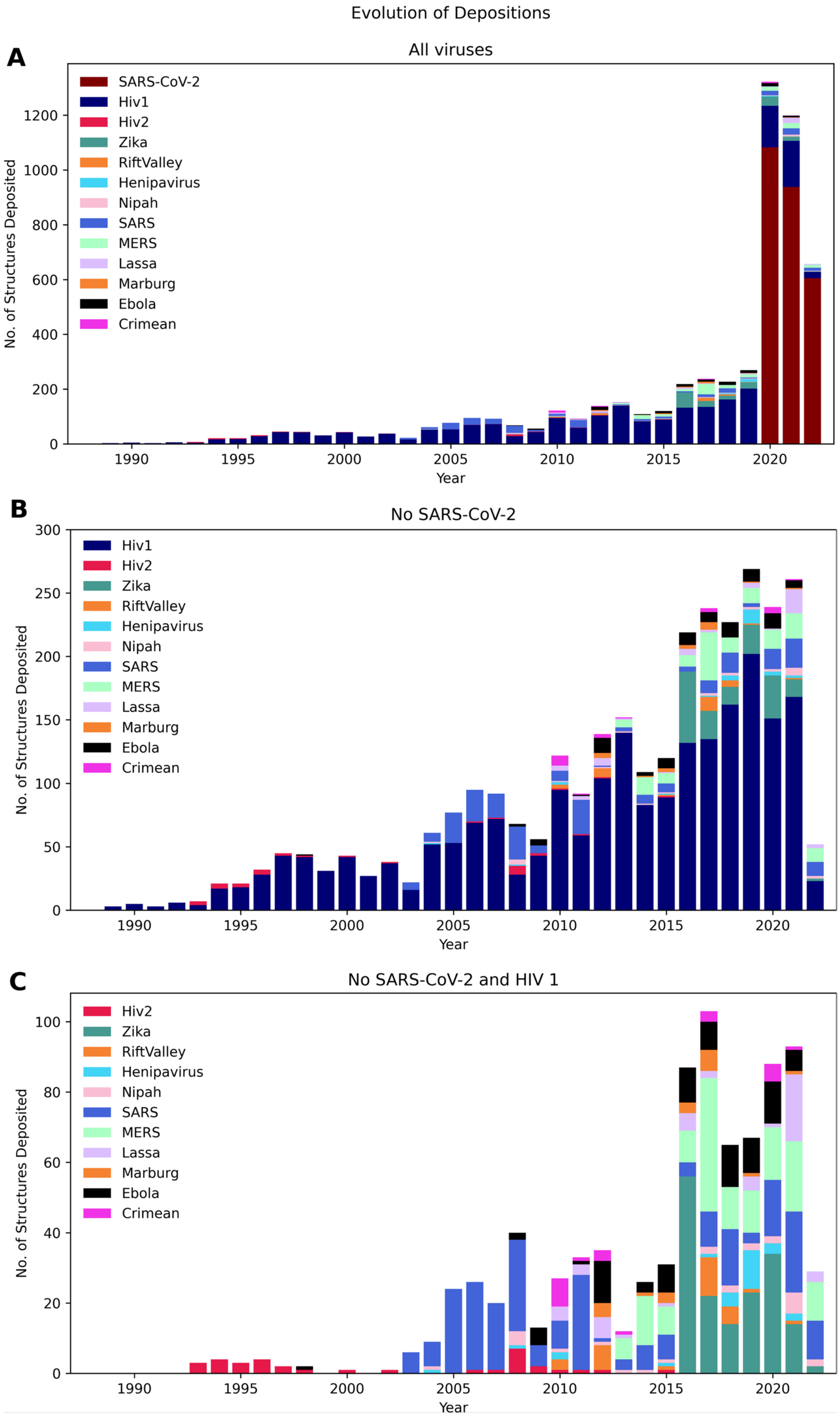

Fig. 1 Time-evolution of the viral structures deposited in the PDB for the selected viruses in this work. (Note the difference in scale; see text for details)

of structures obtained falls sharply to a maximum of 200 entries per year. A clear dominance can be observed in this case for HIV 1, followed by minor spikes in structures from other viruses, namely SARS-CoV-1, MERS, and Zika. Further removal of HIV 1 structures from the plot (Fig. 1C) reduces the number of annual entries to less than 50 entries annually, with the exception of 2016 following the Zika epidemic of 2015/2016. In the absence of SARS-CoV-2 and HIV 1, most structures deposited are from SARS-CoV-1, MERS, Zika, Ebola, LASSA and Crimean virus. Altogether, in the context of the WHO priority list of diseases, the analyzed data indicates that there is a lack of species diversity in the so far obtained viral protein structures. This observation, nonetheless, must be taken considering the timeframe since the establishment of the WHO priority list, originally proposed in 2017 in response to a previous Ebola virus major outbreak. Thus, not enough time might have passed to ensure viral diversity in the PDB depositions. Simultaneously, efforts were made to supply the demand for specific infectious agents in times of crisis, especially SARS-CoV-2. The advances in tools and methodologies (Stollar and Smith 2020), particularly in cryo-EM, along with advances in beamlines, diffractometers and detectors, also account for the outstanding increase in structure depositions during the SARS-COV-2 pandemic. The data presented here might also overlook deposited complexes that include viral peptides and non-viral binders, considering the taxonomy-based approach used to obtain the structural records.

Considering that structure-based drug design (SBDD) or rational drug design are important fields in the discovery of new drug candidates, obtaining novel structures is an essential step for antiviral development. Diseases with epidemic potential like Rift Valley Fever, Henipah/Nipah and Marburg are far behind in terms of available structures to perform such studies, despite not requiring the highest biosafety level for cloning and expressing the proteins of interest. One way to alleviate this issue is to make use of computational methods (e.g. comparative modeling, fold recognition, de novo modeling) or tools like AlphaFold 2 (Jumper et al. 2021) to provide theoretical solutions to the unknown protein structures. These models are frequently much faster to obtain and of acceptable quality. One example is described in a recent report (Narykov et al. 2021) which has shown that computational models of SARS-CoV-2 proteins produced using a combination of comparative modeling and de novo modeling achieved reasonably accurate structures (average root mean squared deviation error of 4.1 Å), while covering 80% of the viral protein sequence (vs 82% from experimental structures). On average, the computational structures were obtained 86 days earlier than the experimental ones. This shows the potential of computational methods in accelerating SBDD projects, especially with novel AI-powered prediction techniques (such as AlphaFold 2) being pushed forward. Despite that, experimentally obtained structures are still of utmost importance, especially when studying the diversity of multiprotein complexes which are commonly formed by viral proteins (Kuhlman and Bradley 2019) and which are harder to predict when using only computational methods. Likewise, AI-based prediction techniques might need additional data to fully encompass the diversity of viral protein structure repertoire (Narykov et al. 2021).

Viral diseases have caused a major impact in human life. Smallpox and HIV are good examples of viruses that have taken the life of millions and millions of people (Nathanson 2016). New viral pandemics or epidemics are expected to occur in the future, possibly from newly emerging or re-emerging viral diseases. For example, if we take the three viruses in our work with the least structures resolved in the PDB, namely Marburg virus, Crimean-Congo hemorrhagic fever virus and Nipah/Henipah virus, we have diseases with, respectively, 24–88%, 40%, and 40–70% case fatality rates. Such high mortality rates are alarming and highlight the importance of better understanding these viral diseases. Current forecasts indicate that deforestation, climate change and the viral diversity in itself are all concurring to promote more frequent viral spillovers of pandemic proportions (Carlson et al. 2022; Pandit et al. 2022). From this perspective, in this report, we demonstrate that there is a need to increase the efforts in viral structures determination in order to increase preparedness for future challenges.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00726-023-03374-2.

Acknowledgements The authors acknowledge Professor Cristina B.C. Bonorino for her guidance and support. All authors are supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Ministério da Educação, Brazil.

Author contributions Conceptualization: RL-B; methodology: MBBG; formal analysis and investigation: MBBG, CP, PB-C; writing—original draft preparation: CP, PB-C, RL-B; writing—review and editing: CP, RL-B; supervision: RL-B. All authors read and approved the final manuscript.

Data availability All data pertaining to this article is available in the Supplementary Information.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Batool M, Ahmad B, Choi S (2019) A structure-based drug discovery paradigm. Int J Mol Sci 20:2783. https://doi.org/10.3390/ijms20112783

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235

Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Ganesan S, Goodsell DS, Ghosh S, Green RK, Guranović V, Guzenko D, Hudson BP, Lawson CL, Liang Y, Lowe R, Namkoong H, Peisach E, Persikova I, Randle C, Rose A, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Tao YP, Voigt M, Westbrook JD, Young JY, Zardecki C, Zhuravleva M (2021) RCSB protein data bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 49:D437–D451. https://doi.org/10.1093/nar/gkaa1038

Carlson CJ, Albery GF, Merow C, Trisos CH, Zipfel CM, Eskew EA, Olival KJ, Ross N, Bansal S (2022) Climate change increases cross-species viral transmission risk. Nature 607:555–562. https://doi.org/10.1038/s41586-022-04788-w

Chatterjee P, Nair P, Chersich M, Terefe Y, Chauhan AS, Quesada F, Simpson G (2021) One health, "Disease X" & the challenge of "Unknown" unknowns. Indian J Med Res 153:264–271. https://doi.org/10.4103/ijmr.IJMR_601_21

Ciotti M, Ciccozzi M, Terrinoni A, Jiang WC, Wang CB, Bernardini S (2020) The COVID-19 pandemic. Crit Rev Clin Lab Sci 57:365–388. https://doi.org/10.1080/10408363.2020.1783198

da Silva SJR, do Nascimento JCF, Germano Mendes RP, Guarines KM, da Silva CTA, da Silva PG, de Magalhães JJF, Vigar JRJ, Silva-Júnior A, Kohl A, Pardee K, Pena L (2022) Two years into the COVID-19 pandemic: lessons learned. ACS Infect Dis. 8:1758–1814. https://doi.org/10.1021/acsinfecdis.2c00204

Engelman A, Cherepanov P (2012) The structural biology of HIV-1: mechanistic and therapeutic insights. Nat Rev Microbiol 10:279–290. https://doi.org/10.1038/nrmicro2747

Hol WGJ (1986) Protein crystallography and computer graphics: toward rational drug design. Angew Chem Int Ed 25:767–778. https://doi.org/10.1002/anie.198607673

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

Kuhlman B, Bradley P (2019) Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 20:681–697. https://doi.org/10.1038/s41580-019-0163-x

Lynch ML, Snell EH, Bowman SEJ (2021) Structural biology in the time of COVID-19: perspectives on methods and milestones. IUCrJ 8:335–341. https://doi.org/10.1107/S2052252521003948

Mandal S, Moudgil M, Mandal SK (2009) Rational drug design. Eur J Pharmacol 625:90–100. https://doi.org/10.1016/j.ejphar.2009.06.065

Mavromoustakos T, Durdagi S, Koukoulitsa C, Simcic M, Papadopoulos MG, Hodoscek M, Grdadolnik SG (2011) Strategies in the rational drug design. Curr Med Chem 18:2517–2530. https://doi.org/10.2174/092986711795933731

Narykov O, Srinivasan S, Korkin D (2021) Computational protein modeling and the next viral pandemic. Nat Methods 18:444–445. https://doi.org/10.1038/s41592-021-01144-0

Nathanson N (2016) The human toll of viral diseases: past plagues and pending pandemics. Viral Pathog 1:3–16. https://doi.org/10.1016/B978-0-12-800964-2.00001-X

Pandit PS, Anthony SJ, Goldstein T, Olival KJ, Doyle MM, Gardner NR, Bird B, Smith W, Wolking D, Gilardi K, Monagin C, Kelly T, Uhart MM, Epstein JH, Machalaba C, Rostal MK, Dawson P, Hagan E, Sullivan A, Li H, Chmura AA, Latinne A, Lange C, O'Rourke T, Olson S, Keatts L, Mendoza AP, Perez A, de Paula CD, Zimmerman D, Valitutto M, LeBreton M, McIver D, Islam A, Duong V, Mouiche M, Shi Z, Mulembakani P, Kumakamba C, Ali M, Kebede N, Tamoufe U, Bel-Nono S, Camara A, Pamungkas J, Coulibaly KJ, Abu-Basha E, Kamau J, Silithammavong S, Desmond J, Hughes T, Shiilegdamba E, Aung O, Karmacharya D, Nziza J, Ndiaye D, Gbakima A, Sajali Z, Wacharapluesadee S, Robles EA, Ssebide B, Suzán G, Aguirre LF, Solorio MR, Dhole TN, Nga NTT, Hitchens PL, Joly DO, Saylors K, Fine A, Murray S, Karesh WB, Daszak P, Mazet JAK, PREDICT Consortium, Johnson CK (2022) Predicting the potential for zoonotic transmission and host associations for novel viruses. Commun Biol 5:844. https://doi.org/10.1038/s42003-022-03797-9

Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr (2013) Computational methods in drug discovery. Pharmacol Rev 66(1):334–395. https://doi.org/10.1124/pr.112.007336

Stollar EJ, Smith DP (2020) Uncovering protein structure. Essays Biochem 64:649–680. https://doi.org/10.1042/EBC20190042

WHO (2023) Prioritizing diseases for research and development in emergency contexts. World Heatlh Organization. https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts. Accessed 22 May 2023

Woolhouse M, Scott F, Hudson Z, Howey R, Chase-Topping M (2012) Human viruses: discovery and emergence. Philos Trans R Soc Lond B Biol Sci 367:2864–2871. https://doi.org/10.1098/rstb.2011.0354

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.