The Covid-19 Pandemic and the Patterns of Nature

Gregory Warr*, Les Hatton†

May 21, 2024

Abstract

This paper addresses broadly the impact that unprecedented levels of scientific discovery can have on the emergent global patterns that we observe in nature. An essentially ubiquitous pattern that is associated with large complex discrete systems is attributable to the Conservation of Hartley-Shannon Information (CoHSI). One of the manifestations of CoHSI in the realm of protein structure is a distinctive equilibrium distribution of protein lengths that is dominated by a power law. Here we examine the manner in which the accelerated pace of novel protein discovery during the Covid-19 pandemic affected this distribution, showing that despite an initial disruption, nevertheless the equilibrium state was reestablished.

1 Introduction

This paper uses a novel approach to study change in the rapidly evolving and globally accessible TrEMBL protein databases available at [1].

The TrEMBL protein databases accumulate the sequenced proteins which result from the efforts of countless teams of researchers around the planet. After more than 25 years of growth, they are already extremely large and still growing rapidly with the most recent release at the time of writing being release 2024_02 dated 27-Mar-2024 of UniProtKB/TrEMBL which contains 248,234,451 sequence entries, comprising 87,367,689,973 amino acids. The proteins vary in length from the shortest A0A0G2JLF7_HUMAN at just 7 amino acids all the way up to a staggering 45,354 amino acids in the longest currently known, A0A5A9P0L4_9TELE.

Clearly these enormous numbers are effectively intractable in terms of identifying local patterns, or even phylogenetically shared

*Department of Biochemistry and Molecular Biology, Medical University of South Carolina, Charleston SC 29425 USA. warrgw@musc.edu

†Faculty of Science, Engineering and Computing, Kingston University, Kingston, UK. lesh@cantab.net

patterns but the discipline of statistical physics over the last 150 years from its origins in the kinetic theory of gases in the hands of the visionary physicist Ludwig Boltzmann, has produced techniques for dealing with such extraordinarily large numbers. Compared with the number of molecules in 1 cubic meter of gas at standard temperature and pressure (2.68×10^{25}), even the TrEMBL databases pale into insignificance. In spite of these vast numbers, the methodology of statistical physics comfortably handles them automatically aggregating any and all local mechanisms to go directly to the equilibrium or most likely distribution. In the case of a gas, the velocities all asymptote to the Maxwell-Boltzmann distribution [2].

By incorporating information theory into this methodology, [3, 4] it was shown that *all* discrete systems (systems composed of countable pieces) sharing only the property that their individual pieces are distinguishable and requiring no other commonality, exhibit patterns dominated by a power-law. These patterns arise from a conservation principle, CoHSI or the Conservation of Hartley-Shannon Information. If we consider the most basic discrete system that consists of ordered strings (components) of coloured beads, then from CoHSI the theoretically predicted length distribution for any discrete system looks like Fig. 1.

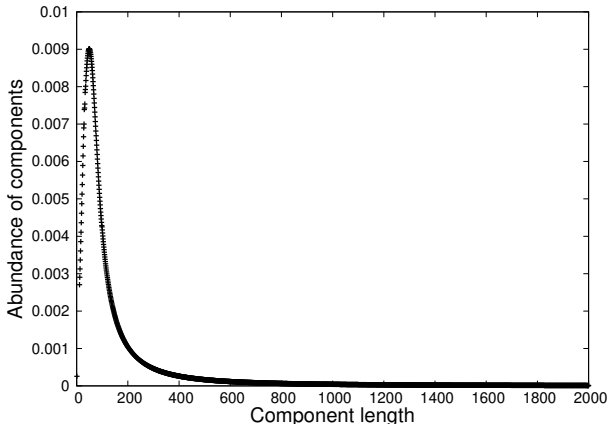

Figure 1: The predicted asymptotic probability distribution function for a set of strings (components) of coloured beads of various lengths with no other property than that the different colours are distinguishable. The distribution shows a sharp unimodal peak transitioning into an extremely precise power-law tail [3].

The presence of this distribution has been identified now in a wide variety of dissimilar discrete systems - lengths of words in texts, number of words in sentences in texts, in large collections of software irrespective of their language or functionality and most notably for the purposes of this paper, in proteins [3]. The prior knowledge

that such an equilibrium distribution of lengths is present in any substantial collection of proteins guides our novel approach allowing us to measure significant departures from that equilibrium state and understand the reasons for any such departures and this we now do.

Proteins are an exemplary discrete system, in that we can consider them as strings of amino acids, the length of a protein being measured by the total number of amino acids. The TrEMBL database[1] represents essentially the totality of our knowledge of the structure and diversity of proteins. One might expect that the distribution of protein lengths would be shaped by natural selection acting on particular structure/function properties, but whereas individual proteins will be subject to natural selection in the normal way, the overall distribution of the properties of proteins results from the complex interactions of many processes; natural selection, genetic drift, random extinctions etc. CoHSI, because of its statistical mechanical framework [3] aggregates all these mechanisms and predicts a scale-free equilibrium outcome that is simply the overwhelmingly most likely state. The theoretically predicted length distribution for any discrete system looks like Fig. 1.

Thus it was predicted (and borne out experimentally) that the length distributions of proteins would show the scale-free distributions implied by CoHSI [3, 5, 4].

Prior to the Covid-19 pandemic we can clearly see the predicted CoHSI distribution in protein lengths, for example in the 2017 TrEMBL release 17-03 (we consider other releases of TrEMBL as this narrative develops) as Figs 2a - 2b. The similarity between the CoHSI prediction Fig. 1 and the data of Fig. 2a is compelling both visually and statistically.

Fig. 2b is a cumulative complementary distribution function [6], a widely used noise-suppressing display of the same data as Fig. 2a. The left hand axis is the number of proteins longer than the size shown on the x-axis. On the left hand side, it is flat corresponding to the sharp rise to the peak of Fig. 1. Reading off this plateau height on the y-axis gives the total number of proteins considered here, (just under 1×10^8 in release 17-03). As we move right and the length increases, fewer and fewer proteins are greater than this length and the data becomes the classic straight line on a log-log scale indicating the presence of the predicted power-law. The mere presence of a straight line is only a necessary condition for a power-law. For greater statistical confidence, a sufficiency test must also be run [7, 8]. Details of this are given in [3] where an emphatic power-law is confirmed.

In essence the above development establishes Fig. 2b as an *equi-*

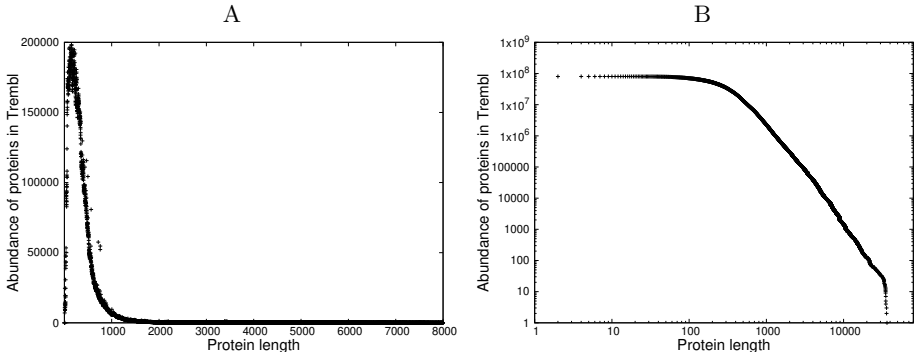

Figure 2: The distribution of lengths of proteins measured in amino acids in TrEMBL release 17-03, A) The distribution as a probability distribution function and B) the distribution as a complementary cumulative distribution function.

librium distribution, an emergent property shared by all discrete systems [4]. In the parlance of statistical mechanics, the mathematical framework behind CoHSI, the biochemical properties of the individual amino acids in the global system of proteins are irrelevant; proteins can be considered as simply consisting of strings of distinguishable amino acids [3]. This pattern of protein lengths shown in 2b is by definition an equilibrium distribution and for such a large distribution, we would normally expect little to disturb this equilibrium. However, there was extraordinary activity in protein discovery, focused on SARS-CoV-2 that took place in 2019 and the years following as a result of the Covid19 pandemic; in the following section we explore the impact of this on the equilibrium distribution of protein lengths.

2 Results and Discussion

2.1 The TrEMBL database 15-07 → 22-02

Having established that there is an equilibrium distribution in protein lengths we can study different versions of TrEMBL as the database grew rapidly in the last few years. Fig. 3 illustrates by taking five releases 15-07, 17-03, 19-04, 21-03 and 22-02. We may first note that the system does indeed closely maintain the equilibrium distribution until the 21-03 distribution where a break suddenly appears in the region of protein lengths of 6,500 to 7,500 amino acids.

Looking at this more closely, the break in Fig. 4a is due to an

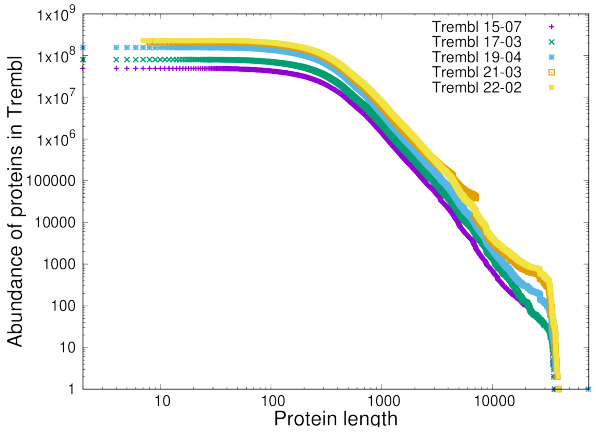

Figure 3: Five recent releases of TrEMBL spanning the Covid-19 pandemic of 2020-2022.

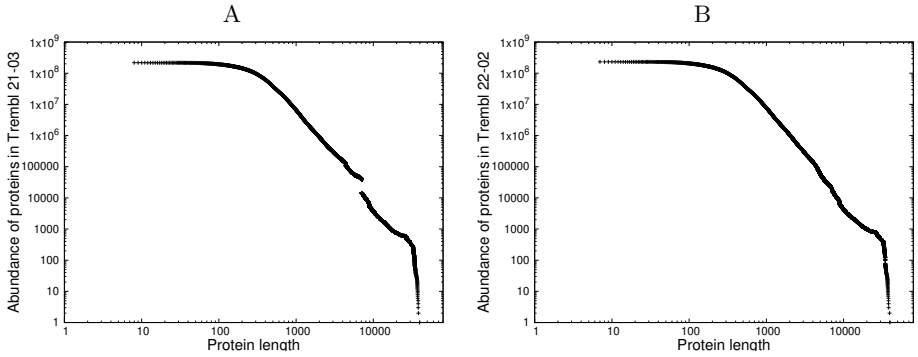

Figure 4: The distribution of lengths of proteins measured in amino acids in TrEMBL, A) Release 21-03 illustrating the clear departure from the equilibrium predicted by CoHSI and due to the uploading of considerable selective work on the SARS-COV-2 virus and B) Release 22-02 12 months later when the equilibrium was essentially restored.

over-abundance (i.e. relative to the equilibrium distribution) of proteins with lengths of approximately 7000 amino acids. Fig. 4a shows that 12 months later the break was already healed and the database resumed its natural growth trend around the equilibrium distribution as defined by CoHSI and exhibited in every other TrEMBL release.

2.2 Covid-19 and the Equilibrium of Protein Lengths

What happened between TrEMBL releases 21-03 and 22-02 to explain 1) why the CoHSI equilibrium was perturbed, and 2) how it

was re-established?

Although the only constraints in CoHSI theory[3] are the total size of the system and its total information content, it is necessary that the system be categorized in a consistent manner. In the case of the TrEMBL database, consistency means that there is no redundancy in the sequence entries. In other words multiple depositions of proteins with identical amino acid sequence and length, and present in the same species (or species equivalent, such as viral strain) are not permitted. While such redundancy of protein entries is eliminated by the curation of the database, curation was relaxed early in the Covid-19 pandemic, when a special portal https://www.ebi.ac.uk/training/events/uniprot-covid-19-website/, now closed, was created by UniProt for the submission of SARS-CoV-2 sequences. Tens of millions of SARS-CoV-2 protein sequences have been uploaded to the protein databases, and we note that the ORF1ab polyprotein of SARS-CoV-2 contains 7096 amino acids. We suggest that the massive uploading of presumptively redundant SARS-CoV-2 sequences resulted in the perturbation of the equilibrium seen in TrEMBL release 21-03. The resumption of normal curation of the database would have eliminated redundancies created by this large volume of identical submissions of SARS-CoV-2 proteins, reestablishing the equilibrium as seen in TrEMBL release 22-02.

Thus while the CoHSI equilibrium as exemplified globally in protein lengths is remarkably stable, at the same time it is sensitive to the consistency of categorization as revealed by the unprecedented number of presumably redundant SARS-CoV-2 sequences that were submitted to TrEMBL early in the Covid-19 pandemic.

2.3 The Covid-19 Pandemic in Perspective

While the Covid-19 pandemic perturbed the equilibrium of protein length distributions, as described above, this resulted from the unprecedented burst of research into the SARS-CoV-2 virus.

However, many other aspects of the Covid-19 pandemic also show power law behaviour, as would be expected from any large, complex discrete system and as predicted by CoHSI theory[3, 4]. Examples of power law distributions can be found early in the course of the pandemic as the infection spread essentially without control and before cases began to reach saturation. Blasius[9] examined the relative size of outbreaks in countries that reported statistics and showed that both the number of infected people and the number of deaths displayed power law distributions. Similar results showing

a power law distribution were reported for Covid-19 fatalities in European countries[10]. Blasius[9] also examined the statistics for SARS-CoV-2 infections and death reported by counties within the United States; these also showed power law distributions.

These results from the Covid-19 pandemic are not unique; the general presence of power laws in epidemics of infectious diseases has been known for some time[11, 12]; for example Rhodes and Anderson[11, 13] showed that both the size and duration of measles epidemics were characterized by power law distributions. It is worth pointing out that early in the response to the Covid-19 pandemic, when specific vaccines were first available, levels of immunization were highly unequal between countries. As would have been predicted, the number of individuals immunized within individual countries was also observed to follow a power law distribution[4].

In conclusion it is reasonable to ask why power laws, as described here in the impacts of the Covid-19 pandemic are essentially ubiquitous in the natural world. As reviewed in detail in [4], power laws are observed in phenomena as diverse as wealth and the frequency of word use, from the size of computer programs to the quantity of alcohol consumed, and from the growth of oyster shells to the size of craters on the moon. Logically, there are only two possibilities. Either there is a single underlying principle that generates power law behaviour in complex discrete systems, or there are many specific local mechanisms that coincidentally generate the same outcomes, i.e. power law distributions, in very different systems. CoHSI theory[3, 4] provides a resolution of these two explanations; complex discrete systems are mechanistically complicated (and often seemingly random) but regardless of any and all mechanisms, distributions dominated by power laws are the essentially inevitable equilibrium state of these systems.

2.4 Ethics

No human subjects, human tissues or animals were used in this research.

2.5 Data Accessibility

This study adheres to the transparency and reproducibility principles espoused by [14, 15, 16, 17, 18, 19] and includes references to all methods and source code necessary to reproduce the results presented. For this study, the methods and source code are included in the wider set of *reproducibility deliverables*, available at `https://`

datadryad.org/stash/share/9nVGYwauP_wdFM84hA6GlC52t4pFircx4NAGl_ukbYA. Each reproducibility deliverable allows all results, tables and diagrams to be reproduced individually for that study, as well as performing verification checks on machine environment, availability of essential open source packages, quality of arithmetic and regression testing of the outputs [20].

2.6 Supplementary Materials

No supplementary materials directly accompany this paper.

2.7 Authors' Contributions

LH performed the analyses, LH and GW developed the arguments, discussed the results and contributed to the text of the manuscript. Both authors gave final approval for publication.

2.8 Competing Interests

The authors declare no competing interests.

2.9 Funding

No institutional or external funding for this research was received by the authors.

3 Acknowledgments

The authors would like to acknowledge the many researchers with whom they have discussed the implications of CoHSI in biological systems over the years, most notably the late Professor Bob Chapman who gave freely of his insights and vast experience.

References

[1] Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research. 2022 11;51(D1):D523–D531. Available from: https://doi.org/10.1093/nar/gkac1052.

[2] Sommerfeld A. Thermodynamics and Statistical Mechanics. Academic Press, New York NY; 1956.

[3] Hatton L, Warr G. Strong evidence of an information theoretical conservation principle linking all discrete systems. RSoc open sci. 2019 11;6(191101).

[4] Hatton L, Warr GW. Exposing Nature's Bias: the Hidden Clockwork behind Society, Life and the Universe. Bluespear Publishing; 2022. Isbn 978-1-908-42204-0.

[5] Hatton L, Warr G. Protein Structure and Evolution: Are They Constrained Globally by a Principle Derived from Information Theory? PLOS ONE. 2015;10:e0125663.

[6] Newman MEJ. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 2006;46:323–351.

[7] Clauset A, Shalizi CR, Newman MEJ. Power-Law Distributions in Empirical Data. SIAM Review. 2009;51(4):661–703. Available from: https://doi.org/10.1137/070710111.

[8] Clauset A. Inference, Models and Simulation for Complex Systems; 2011. Lectures. Available from: http://tuvalu.santafe.edu/\~aaronc/courses/7000/csci7000-001_2011lL2.pdf,accessed24-Jun-2017.

[9] Blasius B. Power-law distribution in the number of confirmed COVID-19 cases. Chaos. 2020;30:093123. Available from: https://pubmed.ncbi.nlm.nih.gov/33003939.

[10] Xenikos DG, Asimakopoulos A. Power-law growth of the COVID-19 fatality incidents in Europe. Infectious Disease Modelling. 2021;6:743–750. Available from: https://www.sciencedirect.com/science/article/pii/S246804272100035X.

[11] Rhodes CJ, Anderson RM. Power laws governing epidemics in isolated populations. Nature. 1996;381:600–602.

[12] Meyer S, Held L. Power-law models for infectious disease spread. The Annals of Applied Statistics. 2014;8(3):1612 – 1639. Available from: https://doi.org/10.1214/14-AOAS743.

[13] Rhodes CJ, Anderson RM. A scaling analysis of measles epidemics in a small population. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 1996;351(1348):1679–1688. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1996.0150.

[14] Popper K. The Logic of Scientific Discovery. Routledge; 1959.

[15] Ziolkowski AM. Further Thoughts on Popperian Geophysics–the Example of Deconvolution. Geophysical Prospecting. 1982;30:p.155–165. Available from: doi:10.1371/journal.pcbi.1003285.

[16] Claerbout JF, Karrenbach M. Electronic documents give reproducibility a new meaning. In: Proc. 62nd Ann. Int. Meeting. Soc. of Exploration Geophysics; 1992. p. 601–604.

[17] Hatton L, Roberts A. How accurate is scientific software ? IEEE Transactions on Software Engineering. 1994;20(10):785–797.

[18] Shahram M, Stodden V, Donoho DL, Maleki A, Rahman I. Reproducible Research in Computational Harmonic Analysis. Computing in Science & Engineering. 2009 jan;11(01):8–18.

[19] Ince DC, Hatton L, Graham-Cumming J. The case for open program code. Nature. 2012 02;482:485–488. Doi:10.1038/nature10836.

[20] Hatton L, Warr G. Full Computational Reproducibility in Biological Science: Methods, Software and a Case Study in Protein Biology. ArXiv. 2016;Available from: `http://arxiv.org/abs/1608.06897[q-bio.QM]`.