Literary Review of Computation Biology

Tai Duc Nguyen
ECES-641
Drexel University
October 26, 2020

This paper provides a quick review of the field of Computational Biology. Through the lenses of the on-going researches on the Covid-19 epidemic, the field's importance to modern societies can be clearly seen and explained.

The National Institute of Health (USA) defined Computational Biology as the science of using biological data to develop algorithms or models in order to understand biological systems and relationships [1]. However, many people have mistaken Computational Biologists for Computer Scientists. This common mistake arises due to the fact that: like Computer Scientists, Computational Biologists do a lot of programming, modeling, producing algorithms to extract and analyze data. However, instead of caring about computing systems and how to make algorithms running on some particular hardware faster, more robust, they care about the relationships between biological systems, the interaction between cells, animals, humans with the surrounding ecosystems. In addition, while Computer Science as a field started in the 1600s with Gottfried Leibniz, who demonstrated a digital mechanical calculator (the Stepped Reckoner) using the binary numeric system [2], Computational Biology only dated back to the 1970s [3], branching off of Bioinformatics - the science of analyzing informatics processes of various biological systems. In fact, researchers in 1982 were sharing information among each other using punch cards. However, by the end of the 1980s, more powerful hardware along with the explosion of mass storage media (propelled by the mp3 format and the invention of the compact disk), allowed researchers to develop algorithms that processed data quickly and inferred relevant relationships. Fast-forward to the 2010s and now 2020, Computational Biology has become one of the foremost frontiers of Biology due to the use of Machine Learning techniques, in particular, Neural Networks. With such potent tools in hand, practitioners of the field are tackling problems in Anatomy, Biomodelling, Genomics, Neuroscience, Pharmacology, Evolutionary Biology, etc. just to name a few. Since it will take a very long time to describe the accomplishments in these subfields, this paper will only focus on Computational Genomics, describing its achievements, importance, and challenges.

Researchers in the Computational Genomics field focus on, as the name suggests, genomes - the genetic material of organisms. In other words, it's the study of life itself. Although it is still a controversy, many scientists believe that if we can fully understand our own genetic code, then we can fully understand life. In the book "Life 3.0: Being Human in the Age of Artificial Intelligence", brilliant Swedish-American cosmologist Max Tegmark defined life as "a process that can retain its complexity and replicate", a.k.a a self-replicating information-processing system whose information (software) determines both its behavior and the blueprints for its hardware. He went further and categorized life into 3 stages: Life 1.0, Life 2.0, and Life 3.0. Organisms (or things) which are in the 1.0 stage have both the hardware and the software(*) evolved rather than designed. Humans are archetypes of the 2.0 stage: "life whose hardware is evolved, but whose software is largely designed." And as the definition goes, Life 3.0 is life where both the hardware and software are largely designed [4]. In order to leap from 2.0 to 3.0, we need to have the capability to design our own genetic code, enabling us to "control our own destiny". Although some people think this leap is inevitable and we should prepare ourselves for it, others think that this ability to modify our own hardware is like "playing God" and should be ethically avoided [5]. Nevertheless, Computational Genomics plays the most crucial part in our journey, for better or for worse, to Life 3.0. The practitioners of this field have achieved major milestones in biology, including:

1. Proposing cellular signaling networks
2. Proposing mechanisms of genome evolution
3. Predict precise locations of all human genes using comparative genomics techniques with several mammalian and vertebrate species
4. Predict conserved genomic regions that are related to early embryonic development
5. Discover potential links between repeated sequence motifs and tissue-specific gene expression
6. Measure regions of genomes that have undergone unusually rapid evolution

The most well-known accomplishment is the "Human Genome Project" (HGP). "HGP was the international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings" [6]. Francis Collins, then director of the National Human Genome Research Institute, this mapping is like a book with multiple uses: "It's a history book - a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease." Even though the entire picture has been revealed, it is such a complex and mysterious mosaic that it still cannot be fully understood after 17 years. In fact, this "book" brings more questions than it answers. One of the most important benefits of HGP was the insight into how our proteins are encoded into our genes and how our "codes" tell for the protein to "fold" like they are. With this understanding, researchers can now understand not only human's proteins but proteins of other organisms, including viruses [7].

In order to combat a certain virus, researchers must have a good understanding of its mechanisms, especially: how it injects its genetic material into the cell; how it travels through different types of medium, including blood and other body fluids; what are its strains and the different types of mutations it can have due to randomness or evolutionary pressure; and what are its weaknesses in terms of switches in its molecular structure (i.e. the membrane, the injector). Hence, many researchers have focused their attention on the mapping of the virus's RNA to its protein structure to understand how it works and how it's made. In early 2020, the SARS-CoV-2 virus has struck the world as one of the most devastating events his human history: infected 44.3 million and killed 1.2 million [8]. In the US alone, there have been 9 million cases and 232 thousand deaths attributed to this virus. Despite continuous efforts to hinder the spread of the virus, the number of infected is still increasing rapidly and the number of death is still ~1000 in the US. Hence, the importance of conducting researches that identify this virus's mechanisms and its genetic mutations among different infected communities becomes ever more paramount.

In the work by Daniele Mercatelli and Federico M. Giorgi from the Dept. of Pharmacy and Biotechnology, University of Bologna, Italy: "Geographic and Genomic Distribution of SARS-CoV-2 Mutations", the researchers have linked genomic mutations with geographical distribution of the virus. In particular, the original (referenced) genome from Wuhan, China displayed a very low mutation rate (7.23 mutations per sample), which allows for the pin-point of its different strains and relating that to its geographical distribution [9]. After aligning the 48,635 SARS-CoV-2 complete genomic sequences from GISAID consortium, there are 353,341 mutation events in total, which results in an average of 7.23 mutations per sequence, a mode of 6 mutations, and very few samples have more than 15 events. This low mutation rate means the Covid-19 viruses do not differ significantly from continent to continent; however, from countries to countries, there is telling distinctions. Looking at the mutations themselves, the authors found that most of the mutations (58.2%) are single-nucleotide polymorphisms (SNPs), with half of those are silent SNPs and a quarter are mutations in the intergenic regions. Deletions and insertions are relatively rare, accounting for a total of less than 1%. Among the SNPs, the most prevalent mutation (worldwide) is a transversion affecting the 23,403rd nucleotide: A>G, defining the so-called G-clade of the virus. This strand is prevalent in Europe, Oceanic, South America, and Africa. This mutation caused a change in the Spike protein, which is responsible for the initial entry of the virus into the cell. The other two major clades are called "S", after the mutation in ORF8 L84S (Ceraolo and Giorgi, 2020), and "V", from the ORF3a:G251V mutation (co-occurring with NSP6:L37F). Geographically, the G and DR clades are mostly present in Europe, while S and GH are prevalent in the Americas. The "L" clade (the original lineage) is represented largely by countries in Asia. Over time, however, the G clade genomes and its derivates, GH and GR, have been become increasingly more common, while clade L, V, and S genomes are gradually disappearing. Hence, with the onset of a vaccine in the next few months, the mutation distribution of the virus will continue to adapt due to selection pressure. With this knowledge in mind, the authors urge researchers to continue monitoring the viral evolution rigorously and develop tools for ad-hoc cooperation around the world.

Methods to closely trace the movement and spread of the virus had been demonstrated with great efficiency by Alexander Crits-Christoph et. al at the University of California, Berkeley, USA in their paper: "Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants" [10]. Since this virus can be found in human feces, researchers have been using Quantitative Reverse Transcription of Polymerase Chain Reaction (`RT-qPCR`) has been applied to municipal wastewater around the world to detect and quantify the presence and abundance of SARS-CoV-2. If the virus is present, then metatranscriptomic sequencing is used to profile its genetic diversity across infected communities. However, in order to get good results, the samples need to be amplified and heavily processed. In addition, RT-qPCR relies on specific PCR primers, which can fail to detect SARS-CoV-2 strains with mutations in the primed sequence [11]. Hence, Alexander Crits-Christoph et. al have used a new technique to analyze the RNA from sewage collected by municipal utility districts in the San Francisco Bay Area. Through their work, they have demonstrated that their methods can distinguish between the genomes of the "local" virus (in the Bay Area) versus those in California and those around the world. The authors described that their technique of sequencing of viral concentrates and RNA extracted directly from wastewater can accurately identify genotypes of viral strains that are clinically detected in a region, and those not yet detected by clinical sequencing. This finding provides a great advantage because it can be quickly used to predict the virus's spread and its mutation before the alien strains showed up clinically.

As new methods of efficient epidemiological surveillance are introduced, Weitao Sun from Tsinghua University, Beijing, China has looked into why this virus is so well-adapted for humans in his paper: "The discovery of gene mutations making SARS-CoV-2 well adapted for humans: host-genome similarity analysis of 2594 genomes from China, the USA, and Europe" [12]. It is a well-known fact that the host's immune system creates evolutionary (selection) pressure upon the virus. However, the virus also exerts pressure back on to its host's immune system, forcing it to adapt. Hence, this so-called "co-evolution" between viruses and their hosts is hypothesized to be one of the driving factors for the evolution of mammalian and human proteomes. Since viruses have to exchange genetic materials with their hosts, similar nucleotide sequences can be found in both subjects. Hence, Sun studied the previous version of SARS-CoV-2, SARS-CoV, and found that the host-genome similarity (HGS) of SARS-CoV-2 has 122-148% higher than that of SARS-CoV, which explains the virus's highly contagious nature. Specifically, the HGS of the protein ORF6 and ORF8 increase greatly in SARS-CoV-2, which represents enhanced ability in suppressing innate immune. Although the exact functionality of ORF6, ORF8, and accessory proteins in SARS-CoV-2 is not well understood, evidence from the virus's genome indicates that they behave similarly to those in SARS-CoV.

Even though a great deal of knowledge was extracted from the genomic data of SARS-CoV-2, many of its proteins' functionalities are unknown. As Mercatelli, Giorgi and Sun have pointed out, the virus is still evolving and the potential for a viral mutation is non-negligible. Hence, Computational Genomics as a tool is needed more now than ever when combatting viruses and understanding diseases such as cancer. In addition to being a necessary tool, this field acts as a medium to gain knowledge about every known biological system, and, as a frontier for people in the field of machine learning/artificial intelligence to push the boundaries from the evolutionary world today to the technological world of tomorrow.

(*) By software, Tegmark meant all the algorithms and knowledge that the organism uses to process information from its senses and decides what to do with it.

References

[1] BISTIC Definition Committee, "NIH WORKING DEFINITION OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY," www.bisti.nih.gov (National Institute of Mental Health, July 17, 2000), https://web.archive.org/web/20120905155331/http://www.bisti.nih.gov/docs/CompuBioDef.pdf.

[2] Fiona Keates, "A Brief History of Computing | The Repository | Royal Society," Royalsociety.org, 2012, http://blogs.royalsociety.org/history-of-science/2012/06/25/history-of-computing/.

[3] Paulien Hogeweg, "The Roots of Bioinformatics in Theoretical Biology," ed. David B. Searls, *PLoS Computational Biology* 7, no. 3 (March 31, 2011): e1002021, https://doi.org/10.1371/journal.pcbi.1002021.

[4] Max Tegmark, *Life 3.0 : Being Human in the Age of Artificial Intelligence* (London: Penguin Books, 2018).

[5] Larry G. Locke, "The Promise of CRISPR for Human Germline Editing and the Perils of 'Playing God,'" *The CRISPR Journal* 3, no. 1 (February 1, 2020): 27–31, https://doi.org/10.1089/crispr.2019.0033.

[6] "The Human Genome Project," Genome.gov, n.d., https://www.genome.gov/human-genome-project/.

[7] "Human Genome Project: Sequencing the Human Genome | Learn Science at Scitable," Nature.com, 2014, https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/.

[8] Worldometer, “Coronavirus Toll Update: Cases & Deaths by Country of Wuhan, China Virus - Worldometer,” Worldometers.info, 2020, https://www.worldometers.info/coronavirus/.

[9] Daniele Mercatelli and Federico M. Giorgi, “Geographic and Genomic Distribution of SARS-CoV-2 Mutations,” *Frontiers in Microbiology* 11 (July 22, 2020), https://doi.org/10.3389/fmicb.2020.01800.

[10] Alexander Crits-Christoph et al., “Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants,” September 14, 2020, https://doi.org/10.1101/2020.09.13.20193805.

[11] Manu Vanaerschot et al., “Identification of a Polymorphism in the N Gene of SARS-CoV-2 That Adversely Impacts Detection by a Widely-Used RT-PCR Assay,” August 26, 2020, https://doi.org/10.1101/2020.08.25.265074.

[12] Weitao Sun, “The Discovery of Gene Mutations Making SARS-CoV-2 Well Adapted for Humans: Host-Genome Similarity Analysis of 2594 Genomes from China, the USA and Europe,” September 3, 2020, https://doi.org/10.1101/2020.09.03.280727.