Engineering 5 (2019) 1155–1158

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
AI for Precision Medicine—Perspective

Information Science Should Take a Lead in Future Biomedical Research

Kenta Nakai

The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan

ARTICLE INFO

Article history:
Received 25 March 2019
Revised 29 June 2019
Accepted 22 July 2019
Available online 20 September 2019

Keywords:
Data science
Artificial intelligence
Next-generation sequencing
DNA
Cancer genome
Single-cell transcriptomics

ABSTRACT

In this commentary, I explain my perspective on the relationship between artificial intelligence (AI)/data science and biomedicine from a long-range retrospective view. The development of modern biomedicine has always been accelerated by the repeated emergence of new technologies. Since all life systems are basically governed by the information in their own DNA, information science has special importance for the study of biomedicine. Unlike in physics, no (or very few) leading laws have been found in biology. Thus, in biology, the "data-to-knowledge" approach is important. AI has historically been applied to biomedicine, and the recent news that an AI-based approach achieved the best performance in an international competition of protein structure prediction may be regarded as another landmark in the field. Similar approaches could contribute to solving problems in genome sequence interpretation, such as identifying cancer-driving mutations in the genome of patients. Recently, the explosive development of next-generation sequencing (NGS) has been producing massive data, and this trend will accelerate. NGS is not only used for "reading" DNA sequences, but also for obtaining various types of information at the single-cell level. These data can be regarded as grid data points in climate simulation. Both data science and AI will become essential for the integrative interpretation/simulation of these data, and will take a leading role in future precision medicine.

1. Modern biology and new technologies

The development of modern biology has always been fueled by the repeated emergence of new technologies. For example, at the end of the 1960s, there were discussions on the decline of molecular biology (i.e., the potential limitations of understanding biological phenomena in terms of the behavior of underlying macromolecules). This was because many scientists had become aware of a certain limitation in the classical approaches—such as bacteriophage-based experiments—at that time [1]. Several pioneers of the field, including Francis Crick himself, then explored challenges in new directions. With the advent of new technology (e.g., recombinant DNA), however, so-called molecular biology has remained in the mainstream of modern biology. As a much more recent example, the rise and explosive development of next-generation sequencing (NGS) technology have changed biology and medicine, not only quantitatively, but also qualitatively [2,3]. NGS will eventually affect society through, for example, changes in social insurance systems. In this commentary, I would like to introduce my thoughts on the future of biomedical research, after briefly reviewing its relationships with data science and artificial intelligence (AI).

2. Information science has special importance in biomedicine

There is no question of the importance of using computers (i.e., devices dealing with "information") in all fields of scientific research. Nevertheless, I would like to emphasize that the use of computers has exceptional importance in the biological/medical sciences because all life systems are basically governed by their own genetic information (DNA). A famous motto from a *New York Times* article about Leroy Hood comes to mind: "Biology is an information science" [4]. Of course, we are still far from a situation in which only theoretical studies on genome DNA sequences are enough to understand biological phenomena. But the relative importance of computational studies will undoubtedly increase in biomedicine; even experimental studies would be greatly aided by robotics and/or AI. To understand complicated biomedical phenomena, such as cancer, we need to consider the systems (i.e., the interactions between many gene products in many conditions and

https://doi.org/10.1016/j.eng.2019.07.023
2095-8099/

cell types). Such efforts would be impossible without the help of computational techniques such as computer simulations.

3. Data science suits biology well

Another important characteristic of biology is that no (or very few) leading laws or principles—equivalent to Newton's laws in physics, for example—have been found in biology thus far. Ernest Rutherford, a famous physicist, once said that "all science is either physics or stamp-collecting" [5]. Biology may have been in his mind as a typical example of "stamp-collecting." Even after a century, the situation has not changed much. It is possible that this feature of biology is inherent, due to the fact that biological systems have evolved in a rather short-sighted manner, analogous to the development of natural languages. If biological systems and natural languages have evolved analogously, the effective methodologies for studying them should also have something in common. Indeed, like the importance of compiling dictionaries for the study of natural languages, it is very important in the fields of biology and medicine to construct databases, which are used for storing and organizing massive data. For example, a leading academic journal, *Nucleic Acids Research* (Oxford University Press), devotes its entire first issue of every year to the topic of databases [6]. As another example, probabilistic modeling methods, such as hidden Markov models (HMMs), have been successfully used in both fields [7]. I believe that these facts endorse the importance of data science in biomedicine.

In fact, modern biology has made great progress as a data-driven science. In the old days, ingenious (small-scale) experiments were performed to prove certain hypotheses; in contrast, nowadays, massive amount of data, which are produced systematically and in an unbiased manner, are processed to find novel knowledge or hypotheses, in an approach that is sometimes called "data to knowledge" (D2K). This is exactly where data science is required; even with our ignorance of backbone principles, our understanding of biomedicine should be deepened to a level that is beneficial enough to our welfare with the help of data science.

4. AI and biomedicine: A retrospective

In computer science, the study of AI (here, I simply define AI as the attempt to make computers more "intelligent," like humans) has a long history that includes a variety of attempts, some of which are closely related to biomedicine. For example, in the early 1970s, a computer program named MYCIN, which was aimed at diagnosing patients with bacterial infectious diseases, had a great impact on society [8]. As another example, in the late 1970s, the MOLGEN project at Stanford University applied knowledge-based problem-solving to several cases, including to design experiments on genetics [9]. I myself chose the topic of applying AI—more specifically, the knowledge/rule-based expert system—for the interpretation of newly determined genome sequences when I was a PhD student. What I actually did was to construct an "if–then"-type expert system for predicting the subcellular localization of proteins from their amino acid sequences [10,11]. The rules were prepared based on the knowledge of various protein sorting signals and sequence features, such as amino acid composition, which are known to be correlated with their subcellular location. The system was named PSORT and was used in the international yeast genome project, among others. Later, we completely updated the system using a machine learning technique (*k*-nearest neighbor algorithm) so that its updating and optimization with frequently updated training data would become much easier [12,13]. It was made available through the Internet, which was still in its infancy at that time. Since then, the PSORT family of predictors has been widely used by molecular biologists. Presently, the mainstream of AI application to biomedicine seems to be occupied with deep learning (see below) but I believe that traditional attempts to use knowledge bases in biomedicine remains to be important. Such studies are now active in the field of the semantic web [14].

5. AI and biomedicine: Recent exciting developments

It has been recognized that waves of enthusiasm for the impact of AI have occurred several times over the years. It seems evident that we are now seeing such a wave, largely caused by the successes of deep learning and related technology [15]. In the field of biology, one milestone might be the recent success of AI in the Critical Assessment of protein Structure Prediction (CASP) competition, which has been held biennially since 1994. In the CASP, participants receive a set of amino acid sequences of proteins without their known folded (three-dimensional (3D)) structures, and submit their predicted 3D structures, which are critically assessed by the organizer. In the latest CASP13, it turned out that the prediction system named AlphaFold, which was developed by the DeepMind team (which is already famous for its success in the traditional game Go) showed the best prediction accuracy [16]. The protein-folding problem is fundamental and has been studied for many years. Thus, this result is quite extraordinary, although it does not mean that the problem itself has been solved completely. Therefore, it seems likely that similar approaches would contribute much to existing problems in DNA sequence interpretation, which should be beneficial for personalized medicine. For example, AI would be helpful for identifying potential disease-related mutations in the genome sequence of each individual. Indeed, a commercial AI-based system (the IBM Watson for Oncology) gives physicians prioritized treatment options based on various kinds of available data. Recently, a concordance study between this AI system and clinical practices for cancer patients in China was published [17]. Such technology would undoubtedly be useful to ① accelerate the personalized diagnosis of a large number of patients, ② facilitate prompt updating of the system in order to keep up with new incoming data, and ③ optimize the system for specific ethnic groups. A next great challenge may be to unify such machine learning approaches with the above-mentioned knowledge-based ones.

6. Modern biomedicine produces massive data through NGS

As I wrote above, all life systems are constituted based on their own information encoded as DNA sequences (i.e., the genome information). Recent progress in NGS technology has enabled the determination of the entire genome of each individual, which is a sequence of about 3.3 billion bases (actually, each individual basically has two genomes, originating from two parents) at a reasonably affordable cost (about 1000 USD or less) [2,18,19] (Fig. 1). To understand what kind of information is written in the genome DNA, NGS can be useful in various ways. ① Since most diseases are related to a defect or variation of the genome, a comparison of genome DNA sequences between disease patients and their healthy counterparts should be useful for spotting which part(s) of the differences is linked to the disease. This kind of approach is called genome-wide association study (GWAS). Once any candidate position (i.e., locus) of DNA and a certain phenotype is found, another technique called DNA editing (through the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas system) might be applied in order to culture cells to confirm the relationship. ② Similarly, an extensive com-

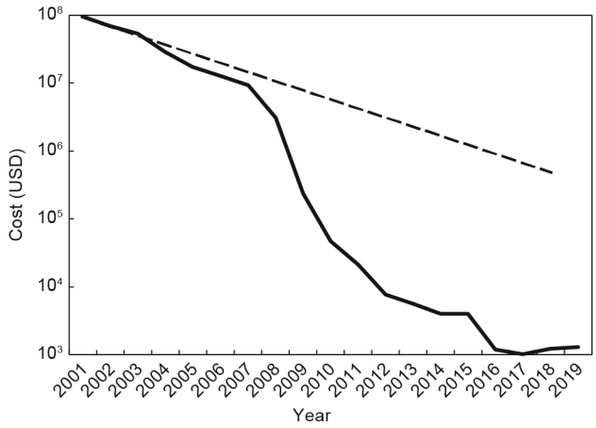

Fig. 1. Trends in the cost of sequencing a human-sized genome, compared with Moore's law. The dotted line representing Moore's law is drawn in a somewhat arbitrary manner [19].

parison of the genome sequences of different species and/or of many individuals should be performed in order to specify which parts of DNA are held in common (i.e., conserved), because these regions are likely to share common functions. It would also be interesting to use such comparisons to find out what kind of evolutionary innovations have taken place via the occurrence of novel change(s) in the genome of a species. For example, since our genome and that of the chimpanzee (and other primates) are quite similar, it is very important to know what the critical differences are between our genomes [20]. ③ Importantly, DNA sequence affects our life not only directly, but also indirectly, through the mechanism known as epigenetics. For example, it is now established that the regions of DNA where genes are actively read are in an exposed structure and are marked with specialized chemical modifications on the DNA itself or on its binding proteins (histones). These marks are used as a kind of cellular memory. Such mechanisms seem to be the key for understanding how a single fertilized egg systematically produces a variety of cells. Interestingly, NGS technology is not only used for "reading" DNA sequences, but also for determining various epigenetic states through techniques such as Chromatin-immunoprecipitation (ChIP) sequencing (ChIP-seq) [21] and Hi-C [22]. Recently, it has even become possible to obtain such data from individual cells (through single-cell sequencing/epigenomics), enabling the precise tracing of the entire development of some simpler organisms on the cellular level [23]. Such a single-cell technology would also be useful in understanding the heterogeneity of cancer cells: how a novel somatic mutation that can boost tumor growth occurs within a population of tumor cells; how a subpopulation of cells with such a mutation proliferates with the progression of tumor stages; and how some of the cells acquire the ability to circulate in body fluids, leading to the spread of cancer to parts of the body distant from its origin (i.e., metastasis) [24]. Indeed, it has been revealed that a trace of fragmented DNA originating from tumor cells circulates in the blood in even relatively early stages of cancer. The technology to detect such DNA (cell-free DNA (cfDNA)) for the prognosis of patients, which is known as liquid biopsy, is set to revolutionize early cancer detection [25]. ④ DNA sequencing is applied not only to purified DNA samples, but also to mixtures of DNA—that is, to DNA originating from a number of species (metagenomics). One typical example is the metagenome sequencing of gut bacteria, from which we can estimate the rough composition of bacteria in the gut. Such information would be quite valuable for understanding human health, because gut bacteria are known to influence our health in various ways by interacting with the body through various metabolites (chemical compounds) and so forth [26]. Thus, in combination with metabolome data, which are systematically obtained using high-throughput mass spectrometers, we can obtain a more precise portfolio of our health status. In summary, NGS can be used in various ways in biomedicine, and all of these efforts will inevitably be expanded to produce a veritable avalanche of data (Fig. 2) [27]. The pace of NGS performance improvement even exceeds that of Moore's law (Fig. 1). This situation must be addressed using data science and AI—in fact, these technologies should lead biomedicine rather than just help to address its issues.

7. Conclusion

When the Human Genome Project was underway about 20 years ago, I heard an interesting analogy between biology and weather forecasting:[†] In our childhood, weather forecasting was done by experienced professionals, but their forecasting was not particularly reliable. Nowadays, a combination of data (e.g., temperature, humidity, and air pressure) is obtained at many grid points and fed into supercomputers. Consequently, the forecasting based on these simulation results has become more reliable. In a similar way, the combination of precise data—such as the variety of NGS data introduced above—that has been measured at a great number of points (e.g., individual cells) will be used to computationally predict various things (e.g., the potential risk of an individual having a disease within the next ten years). Such approaches are currently mentioned in the context of multi-omics and/or precision medicine. Both data science and AI will become essential for the integrative interpretation and simulation of these data. These technologies will indicate what kind of additional information is necessary, and what kind of experiments are needed in order to prove the generated hypotheses. Therefore, the next ten years should become even more exciting for biomedicine.

Acknowledgements

I thank Dr. Le Zhang for inviting me to write this article, and Dr. Ashwini Patil and Dr. Sung-Joon Park for helping me to polish it. This work was partly supported by JSPS KAKENHI (17K00397).

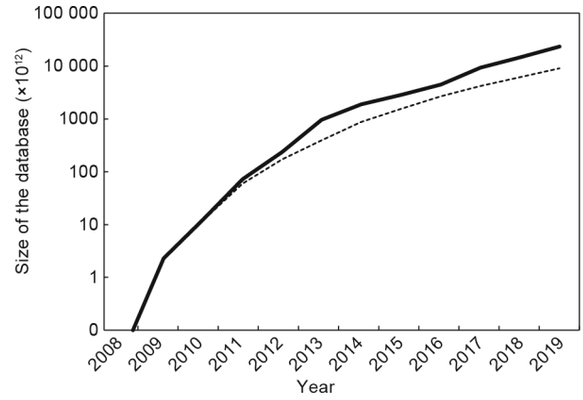

Fig. 2. The amazing growth of NGS data stored in a public database (Sequence Read Archive (SRA) database at the National Center for Biotechnology Information (NCBI), National Institutes of Health (NIH), USA). The *y*-axis shows the size of the database in a logarithmic scale. The solid line shows the total bases, and the dotted line shows the open-access bases (i.e., data downloadable without any restrictions). As of June 2019, the SRA contains more than 2.9×10^{16} bases in total [27].

[†] Personal communication with Masaru Tomita.

References

[1] Stent GS. That was the molecular biology that was. Science 1968;160(3826):390–5.
[2] Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet 2014;30(9):418–26.
[3] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17(6):333–51.
[4] Pollack A. Scientist at work: Leroy Hood; a biotech superstar looks at the bigger picture [Internet]. New York: The New York Times Company; c2019 [cited 2019 Aug 1]. Available from: https://www.nytimes.com/2001/04/17/science/scientist-at-work-leroy-hood-a-biotech-superstar-looks-at-the-bigger-picture.html.
[5] Birks JB, Segrè E. Rutherford at Manchester. Phys Today 1963;16(12):71.
[6] Rigden DJ, Fernández XM. The 26th Annual Nucleic Acids Research Database Issue and Molecular Biology Database collection. Nucleic Acids Res 2019;47:D1–7.
[7] Vijayabaskar MS. Introduction to hidden Markov models and its applications in biology. Methods Mol Biol 2017;1552:1–12.
[8] Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975;8(4):303–20.
[9] Stefik MJ, Martin N. A review of knowledge based problem solving as a basis for a genetics experiment designing system. Stanford: Computer Science Department, Stanford University; 1977.
[10] Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria. Proteins 1991;11(2):95–110.
[11] Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 1992;14(4):897–911.
[12] Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999;24(1):34–6.
[13] Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res 2007;35(Suppl 2):W585–7.
[14] Chen H, Yu T, Chen JY. Semantic web meets integrative biology: a survey. Brief Bioinform 2013;14(1):109–25.
[15] Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nat Biotechnol 2018;36(9):829–38.
[16] AlQuraishi M. AlphaFold at CASP13. Bioinformatics 2019:btz422.
[17] Zhou N, Zhang CT, Lv HY, Hao CX, Li TJ, Zhu JJ, et al. Concordance study between IBM Watson for Oncology and clinical practice for patients with cancer in China. Oncologist 2019;24(6):812–9.
[18] Park ST, Kim J. Trends in next-generation sequencing and a new era for whole genome sequencing. Int Neurourol J 2016;20(Suppl 2):S76–83.
[19] DNA sequencing costs: data [Internet]. Bethesda: National Human Genome Research Institute; [cited 2019 Aug 1]. Available from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.
[20] Kuhlwilm M, de Manuel M, Nater A, Greminger MP, Krützen M, Marques-Bonet T. Evolution and demography of the great apes. Curr Opin Genet Dev 2016;41:124–9.
[21] Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform 2017;18(2):279–90.
[22] Eagen KP. Principles of chromosome architecture revealed by Hi-C. Trends Biochem Sci 2018;43(6):469–78.
[23] Marioni JC, Arendt D. How single-cell genomics is changing evolutionary and developmental biology. Annu Rev Cell Dev Biol 2017;33(1):537–53.
[24] Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer 2017;17(9):557–69.
[25] Hofman P, Heeke S, Alix-Panabières C, Pantel K. Liquid biopsy in the era of immune-oncology. Is it ready for prime-time use for cancer patients? Ann Oncol 2019;30(9):1448–59.
[26] Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 2017;243:16–24.
[27] SRA database growth [Internet]. Bethesda: National Center for Biotechnology Information; [cited 2019 Aug 1]. Available from: https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/.