Contents lists available at ScienceDirect

Engineering

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

Research
AI for Precision Medicine—Perspective

信息科学应引领未来的生物医学研究

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

关键词
数据科学
人工智能
下一代测序
DNA
癌症基因组
单细胞转录组学

摘要

笔者从长期回顾的角度阐述了对人工智能（AI）/数据科学与生物医学之间关系的看法。随着新技术的不断出现，现代生物医学的发展持续加速。由于所有生命系统基本上都受其自身DNA中信息的支配，因此信息科学对生物医学的研究具有特别重要的意义。与物理学不同，在生物学中没有发现（或很少有）主导定律。因此，在生物学中，“数据到知识”方法很重要。人工智能在历史上一直应用于生物医学，最近的新闻表明，基于人工智能的方法在国际蛋白质结构预测竞争中获得了最佳性能，这可能被视为该领域的另一个里程碑。类似的方法可能有助于解决基因组序列解释中的问题，如确定患者基因组中的癌症驱动突变。最近，新一代测序（NGS）的爆炸性发展已产生大量数据，并且这种趋势将加速。NGS不仅用于“读取”DNA序列，而且还用于在单细胞水平上获得各种类型的信息。这些数据可以视为气候模拟中的网格数据点。数据科学和人工智能对于这些数据的综合解释/模拟都将变得至关重要，并将在未来的精密医学中起主导作用。

1. 现代生物学和新技术

新技术的不断涌现一直推动着现代生物学的发展。例如，在20世纪60年代末，曾有关于分子生物学衰落的讨论（即从潜在大分子行为的角度理解生物现象的潜在局限）。这是因为当时许多科学家已经意识到传统方法（如基于噬菌体的实验）存在一定的局限性[1]。该领域的几位先驱，包括Francis Crick本人，随后向新方向的挑战发起探索。然而，随着新技术（如重组DNA）的出现，所谓的分子生物学仍然是现代生物学的主流。更近的例子是，新一代测序（NGS）技术的兴起和暴发式发展不仅在数量上，而且在质量上改变了生物学和医学[2–3]。NGS最终将通过社会保险制度的变化等方式影响社会。在这篇评论中，笔者想在简要回顾生物医学研究与数据科学和人工智能（AI）的关系之后，介绍笔者对未来生物医学研究的看法。

2. 信息科学在生物医学中具有特殊的重要性

毫无疑问，使用计算机在科学研究的各个领域都很重要（如处理“信息”的设备）。然而笔者想强调，计算机的使用在生物（医学科学）中具有特殊的重要性，因为所有的生命系统基本上都由它们自身的遗传信息（DNA）控制。《纽约时报》中一篇关于Leroy Hood的文章中有一句名言:“生物学是一门信息科学。”[4]当然，我们还远远没有达到只对基因组DNA序列进行理论研究就能理解生物学现象的地步。但计算研究的相对重要性无疑将在生物医学领域得到提高，即使是实验研究也

英文原文: Engineering 2019, 5(6): 1155–1158
引用本文: Kenta Nakai. Information Science Should Take a Lead in Future Biomedical Research. *Engineering*, https://doi.org/10.1016/j.eng.2019.07.023Research

会得到机器人和（或）人工智能的极大帮助。要了解复杂的生物医学现象，如癌症，我们就需要考虑系统（即在许多条件下，诸多基因产物与细胞类型之间的相互作用）。如果没有计算机模拟等计算技术的帮助，这样的努力是不可能得到结果的。

3. 数据科学很适合生物学

生物学的另一个重要特征是，迄今为止，在生物学上还没有发现（或很少）与牛顿定律等价的主要定律或原理。著名物理学家Ernest Rutherford曾说“所有的科学要么是物理，要么是集邮”[5]。生物学或许是他心中一个“集邮”的典型例子。即使过了一个世纪，这种情况也没有太大改变。生物学的这一特征可能是与生俱来的，因为生物系统是以一种相当短视的方式进化的，类似于自然语言的发展。如果生物系统和自然语言的进化类似，那么研究它们的有效方法也应该有一些共同之处。的确，就像编纂字典对自然语言研究的重要性一样，在生物学和医学领域，建立用于储存和组织大量数据的数据库也非常重要。例如，核心学术期刊*Nucleic Acids Research*（牛津大学出版社出版）每年的第一期都以数据库为主题[6]；另一个例子是，诸如隐马尔可夫模型（HMM）等概率建模方法已经在这两个领域得到了成功的应用[7]。笔者相信这些事实证明了数据科学在生物医学中的重要性。

实际上，现代生物学作为一门数据驱动科学已经取得了很大的进步。在过去，人们通过巧妙的（小规模的）实验来证明某些假设；与此相反，如今通过处理大量系统化产生的无偏数据可以得到新的知识或假设，这种方法有时被称为“从数据到知识”（D2K）。这正是需要数据科学的地方，即使不知道基本原理，我们对生物医学的理解也应该在数据科学的帮助下加深至足以造福人类的程度。

4. 人工智能和生物医学——回顾

在计算机科学中，对人工智能的研究（这里，笔者只是将人工智能定义为尝试使计算机像人类一样更“智能化”）有着悠久的历史，包括各种各样的尝试，其中一些与生物医学密切相关。例如，在20世纪70年代初，一个名为MYCIN的诊断细菌感染性疾病的计算机程序对社会产生了巨大的影响[8]；另一个例子是，在20世纪70年代末，斯坦福大学的MOLGEN项目将基于知识的问题解决方法应用于多个案例，包括设计遗传学实验[9]。当笔者还是一个博士生时，选择了应用AI的主题——更具体地说，是基于知识/规则的专家系统——来解释新确定的基因组序列。实际上，笔者构建了一个“if-then”类型的专家系统，用于从氨基酸序列预测蛋白质的亚细胞定位[10,11]。这些规则是根据已知与亚细胞位置相关的各种蛋白质分类信号和序列特征（如氨基酸组成）来制定的。该系统被命名为PSORT，并用于国际酵母基因组计划。后来，我们利用机器学习技术（*k*最近邻算法）全面升级了系统，使其更容易在频繁更新的训练数据下完成更新和优化[12,13]。它是通过互联网运行的，当时互联网还处于起步阶段。此后，预测因子PSORT家族得到了分子生物学家的广泛应用。目前，人工智能应用于生物医学的主流似乎是深度学习（见下文），但笔者认为传统的尝试在生物医学中使用知识库仍然很重要。这样的研究现在活跃在语义网领域[14]。

5. 人工智能与生物医学——近期激动人心的发展

近年来，人工智能的影响已几度引发人们的研究热情。很明显，我们现在看到的这种浪潮，很大程度上是由深度学习和相关技术的成功引起的[15]。在生物学领域，一个里程碑可能是人工智能最近在蛋白质结构预测的关键评估（CASP）比赛中的获胜，该比赛自1994年以来每年举行一次。在CASP中，参赛者得到一组折叠[三维（3D）]结构未知的氨基酸序列的蛋白质，并提交他们预测的3D结构，由组织者严格评审。在最近的第十三届CASP中，由DeepMind团队（该团队因其在传统围棋游戏中的成功而闻名）开发的AlphaFold预测系统显示出了最好的预测精度[16]。蛋白质折叠这一基本问题已经被研究了很多年，所以这一结果的意义非凡，尽管它并不意味着问题本身已经完全解决。因此，类似的方法可能会对解决DNA序列解释中存在的问题很有用，这应该有利于个性化医疗。例如，人工智能可能有助于识别每个个体的基因组序列中潜在的与疾病相关的突变。事实上，一个商业化的基于人工智能的系统（the IBM Watson for Oncology）根据各种可用数据为医生提供优先治疗方案。最近，有一项针对中国癌症患者的人工智能系统与临床实践的一致性研究被发表[17]。这种技术毫无疑问有助于：①加速对大量患者的个性化诊

断；②及时更新系统以使其与新传入的数据相匹配；③优化针对特定族群的系统。下一个巨大的挑战可能是将这类机器学习方法与上述知识型方法相结合。

6. 现代生物医学通过 NGS 产生大量数据

正如笔者上面提到的，一切生命系统都是基于它们被编码成DNA序列的信息（也就是基因组信息）而构成的。NGS技术的最新进展使得以合理的成本（约1000 USD或更少）测定每个个体的整个基因组成为可能，这是一个大约3.3×10^9个碱基的序列（实际上，每个个体基本上有两个来自双亲的基因组）[2,18,19]（图1）。NGS在很多方面对于了解基因组DNA中包含的信息是很有用的：①由于大多数疾病都与基因组的缺陷或变异有关，因此将患者和健康人的基因组DNA序列进行比较，应该有助于确定哪些部分的差异与疾病有关。这种方法被称为全基因关联研究（GWAS）。一旦发现DNA的任何候选位置（即所在地）和某种表型，就可以采用另一种被称为DNA编辑的技术[通过规律成簇间隔短回文重复（CRISPR）/Cas系统]来培养细胞以确认这种关系。②与此类似，应该对不同物种和（或）许多个体的基因组序列进行广泛的比较，以确定DNA的哪些部分是相同的（即保守的），因为这些区域可能有相同的功能。同样有趣的是，利用这样的比较可以弄清一个物种基因组的新变化将引发什么样的进化创新。例如，由于人类基因组和黑猩猩（以及其他灵长类动物）的基因组非常相似，因此了解人类基因组之间的关键差异是非常重要的[20]。③重要的是，DNA序列通过表观遗传学机制直接和间接地影响着我们的生活。例如，现在已经证实，基因读取活跃的DNA区域处于暴露的结构中，并且在DNA本身或其结合蛋白（组蛋白）上标记有特殊的化学修饰。这些标记被用作一种细胞记忆。这些机制似乎是理解单个受精卵如何系统地产生各种细胞的关键。有趣的是，NGS技术不仅用于“读取”DNA序列，还可以通过染色质免疫沉淀测序（ChIP-seq）[21]和Hi-C [22]等技术来确定各种表观遗传状态。最近，甚至有可能从单个细胞（通过单细胞测序/表观基因组学）获得这样的数据，从而能够在细胞水平上精确追踪一些更简单的生物体的整个发育过程[23]。这种单细胞技术也有助于理解癌细胞的异质性：一种能够促进肿瘤生长的新的体细胞突变如何在肿瘤细胞群中发生；具有这种突变的细胞亚群如何随着肿瘤的生长而增殖；以及一些细胞是如何获得在体液中循环的能力，从而导致癌症扩散到远离其起源的身体部位（即转移）[24]。事实上，即使在癌症的相对早期阶段，血液循环中也有来自肿瘤细胞的DNA碎片。以预测患者为目的而对这种DNA[无细胞DNA（cfDNA）]进行检测的技术被称为液体活检，它将彻底改变早期癌症检测[25]。④DNA测序不仅适用于纯化DNA样本，还适用于混合DNA，即来自多个物种的DNA（宏基因组）。一个典型的例子是肠道细菌的宏基因组测序，由此我们可以估计肠道细菌的大致组成。众所周知，肠道细菌通过各种代谢产物（化合物）与人体发生作用，以多种方式影响人类的健康，所以这些信息对于了解人类健康是非常有价值的[26]。因此，结合使用高通量质谱仪系统获得的代谢组数据，我们可以获得更精确的健康状况组合。综上所述，NGS可以运用到生物医学的多个方面，人们将持续努力，以产生大量真实的数据（图2）[27]。NGS性能提高的速度甚至超过了摩尔定律（图1）。这种情况必须通过数据科学和人工智能来解决——事实上，这些技术应该引领生物医学，而不仅仅是帮助其解决问题。

7. 结论

大约20年前，当人类基因组计划启动时，笔者耳闻过生物学与天气预报之间一个有趣的类比[†]：在我们的童年时代，天气预报是由经验丰富的专业人士来完成的，

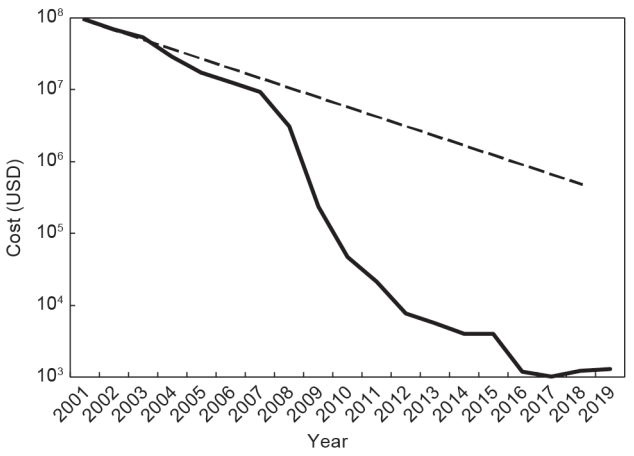

图1. 与摩尔定律相比，人类基因组测序成本的变化趋势。虚线代表摩尔定律，其绘制具有一定随机性[19]。

[†] Personal communication with Masaru Tomita.

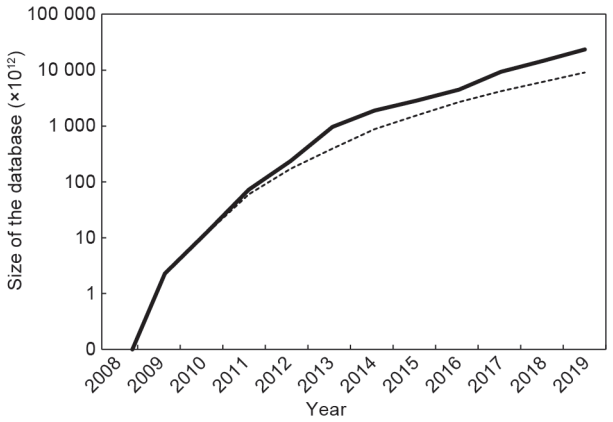

图2. 公共数据库中的NGS数据以惊人的速度增长[美国国立卫生研究院（NIH），国家生物技术信息中心（NCBI）的序列读取存档（SRA）数据库]。*Y*轴以对数刻度表示数据库的规模。实线代表总库数，虚线代表开放的库（即资料下载不受任何限制）。截至2019年6月，SRA总共拥有2.9×10^{16}个库[27]。

但他们的预报并不十分可靠。如今，数据的组合（如温度、湿度和气压）可以在多个网格点获得，并输入超级计算机中。因此，基于这些模拟结果的预测变得更加准确。与之类似，在大量点（如单个细胞）测得的精确数据的组合（如上面介绍的各类NGS数据）将用于计算预测各种事情（如个人在未来10年内患病的潜在风险）。这些方法目前在多组学和（或）精准医疗的背景下被提到。数据科学和人工智能对于这些数据的综合解释和模拟都将变得至关重要。这些技术将表明需要什么样的附加信息，以及什么样的实验来证明生成的假设。因此，未来10年对于生物医学来说将会更加激动人心。

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).

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 nextgeneration 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-thebigger- 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. Computerbased 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/.