Algorithmic biology unleashed

Hallam Stevens

Accepted version

Published version: https://www.science.org/doi/10.1126/science.abg5266

The development of powerful algorithms was necessary for processing, analyzing, and attempting to make sense of the vast amounts of data the Human Genome Project produced. The "assembly" of the genome itself was a process made possible by algorithms (1). Algorithms for searching, matching, and pattern-finding constituted a bridge between the "raw" data of the genome and the practical application of genomic knowledge – they linked DNA to the causes of disease or to evolutionary markers.

Since 2001, algorithms for matching, indexing, and searching DNA and RNA sequences have become ever more central to the way biology is practiced. The Moore's Law-like growth of next-generation sequencing, genome-wide association studies, and personalized genomics have generated ever-increasing troves of data to feed ever-faster algorithms. We can trace a genealogy of algorithms from FASTA to BLAST to SOAP3 that have reduced these data to biomedical knowledge (2). Genomics has been driven by accelerating ways to find patterns in large data sets.

These data-hungry algorithms now not only consume DNA sequence, but also transcriptional, proteomic, metabolomic, and other forms of biological data. This raises new possibilities and new challenges for biomedicine. The most ambitious of these programs have begun mobilizing machine learning to make predictions about our future health, our future offspring, and our future selves (3,4). They imagine instructing us on what to eat, what medicines to take, how to work out, and who to partner with. Algorithms now promise us healthier, happier, and longer ways of living.

Such algorithms have also begun to look towards combining biological data with other forms of "personal data" -- where we go, what we buy, who we associate with, what we "like." In so doing, they draw biomedicine more closely into the orbits of the global Internet giants that are sucking up and attempting to monetize these various forms of data. Biological data risks becoming another kind of data to be bought, aggregated, and used to sell things to us (5).

The hopes for the HGP have long been tempered by fears that the genome could reveal too much about ourselves, exposing us to new forms of discrimination, social division, or control. A predictive and personalized biology is the ultimate DNA dream. But the emerging data-driven biology presents significant challenges for privacy, data ownership, and

algorithmic bias (6,7,8) that must be addressed if genomics is to avoid becoming a handmaiden of "surveillance capitalism" (9).

References

(1) W. J. Kent, D. Haussler, UCSC-CRC-00-17 (27 December 2000). Available at: http://users.soe.ucsc.edu/~learithe/browser/goldenPath/algo.html

(2) H. Stevens, *Perspectives on Science* 19(3), 263-299 (2011).

(3) L. Koumakis, *Computational and structural biotechnology journal* 18, 1466-1473 (2020). https://doi.org/10.1016/j.csbj.2020.06.017

(4) E. Pennisi, *Science*, **357** (6346) 25 (7 July 2017).

(5) E. Ravenscraft (2019), *New York Times* (12 June 2019) https://www.nytimes.com/2019/06/12/smarter-living/how-to-protect-your-dna-data.html

(6) D.Grishin, K. Obbad, G. M. Church, *Nature Biotechnology* **37**: 1115-1117 (2019).

(7) L. Bonomi, Y. Huang, L. Ohno-Machado, *Nature* **52**: 646-654 (2020).

(8) K. Ferryman, M. Pitcan, *Fairness in precision medicine.* (Data & Society, 2019). https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf

(9) S. Zuboff, *The age of surveillance capitalism: the fight for a human future at the new frontier* (Profile Books, 2019).