Published in final edited form as:

Nat Cell Biol. 2024 January ; 26(1): 5–7. doi:10.1038/s41556-023-01286-7.

Bringing computation to biology by bridging the last mile

Anne E. Carpenter,

Shantanu Singh

Imaging Platform, Broad Institute of Harvard and MIT, Cambridge, MA, USA.

Abstract

Advances in technology dramatically accelerate biology research, with computation being a standout example. Typically, adapting a new technology follows stages from method creation, via proof-of-concept application to biology, to the development of usable tools. Creating user-friendly software to bridge computer science and biology is a crucial step, yielding high returns on investment and driving biological discoveries. However, we need dedicated resources and a shift in the academic reward system to harness the full potential of computer science in biology.

In many professions, the everyday course of work remains relatively unchanged over decades. But as scientists, we have the distinct pleasure of looking at our advisor's PhD thesis and thinking, "I could replicate that work in a week." We see a *Nature* paper from the 1980s and think to ourselves how much easier it was to publish in top journals back then – but in reality, advances in technology have paved the way, making experiments ever better, faster and cheaper than in the 'old days'. In fact, the key development that accelerates a field is often not just the invention of a new technology but instead the development of a convenient technology.

Sequencing technology provides a jaw-dropping example of new and convenient instrumentation driving an increase in scale. The first gene (460 nucleotides encoding the coat protein of the bacteriophage MS2) was painstakingly sequenced in 1972 and earned a *Nature* paper[1]. Twenty-nine years later, a draft of the first human genome was finished; more than six-million-fold larger, it also warranted publication in *Nature*[2] and in *Science*[3]. Fourteen years after that, the 1000 Genomes Project announced having sequenced genomes from 2,504 people[4]. Three years later, in 2018, the 100,000 Genomes Project hit its target.

Technology-driven acceleration in biology has been equally dramatic in the application of computation. Students today may not realize that there was a time before the existence of software for carrying out BLAST searches, when scientists aligned gene sequences by eye; although the creation of new alignment algorithms was key, the development of user-friendly tools accelerated discovery. Future students will be shocked to learn that a time existed during which we did not have a protein structure prediction available for nearly all human genes[5], a feat recently accomplished in draft form by Alphafold. In fact, early

anne@broadinstitute.org; shsingh@broadinstitute.org.

Competing interests

The authors declare no competing interests.

structural biologists did not even use software; instead, they applied their expertise (and spatial intuition) to translate spots on X-ray crystallography films into protein structures. Now, convenient tools make it possible to solve and visualize structures.

Turning to cell biology and computation, I (A.E.C.) spent hundreds of hours during my PhD training using a digital ruler to manually measure blobs of chromatin in microscopy images, in order to determine the impact of transcriptional activation on its large-scale structure. Visual screens of thousands of samples were commonplace (and tedious) in the 2000s. Now, high-quality open-source software (such as my own invention, CellProfiler) has turned theoretical computational approaches into user-friendly software that can readily measure nearly any phenotype in millions of images using cloud resources and little hands-on time. For me (S.S.), I used to access graphical processing units for deep-learning bioimage analysis via complex local computer configurations. Now, cloud services such as Google Colab provide instant access to free graphical processing units remotely. This of course has not translated into massive layoffs of research assistants or graduate students around the world; instead, scientists use software to automate tedious work and to increase the scale of their experiments, beyond what was conceivable a decade before.

It is not just that experiments are faster and cheaper and thus more scalable – the very nature of experiments can change with technological advances. For example, I wrote CellProfiler with the goal of helping biologists to measure whatever individual phenotype they cared about in their experiment, but my lab's focus is now dedicated to image-based profiling using Cell Painting, which we use to measure thousands of cell morphology features from fluorescence images simultaneously, and then use the patterns in these data to cluster genes or chemical compounds on the basis of their functional impact. Already, image-based profiling has led to the discovery of previously unknown gene functions and to several potential therapeutics entering clinical trials.

Having seen many fields within biology benefit from computational techniques, we see a recurring pattern of stages in this process from theory to practice (Fig. 1).

Method creation is the first stage. Usually occurring entirely within the computational field, with no application to biology, a new algorithm or machine-learning technique is devised. Some computer scientists focus heavily on the theoretical, so initial papers describing the advancement may have no evidence of practical application to a real-world task, biological or otherwise. Eventually, researchers test new methods for actual tasks, often using well-known benchmarks such as ImageNet (for classifying images that contain everyday objects) or GLUE (for tasks involving the understanding of language). The past decade has brought an explosion of literature in the machine-learning world, in part because of massive funding from much better-resourced industries such as social media, advertising, and finance. It has become impossible to stay current with all that is developing – to the point that most of us resort to machine-learning-powered tools to direct us to the most interesting papers! In theory, new computational methods might be created directly in the service of biological problems, given their distinctive data structures, assumptions and tasks. Still, with some notable exceptions such as UNet for image segmentation[6], the flow of developments tends to primarily be one directional, from computer science to biology.

Proof-of-concept application to biology comes next, whereby a method from the computer science literature is applied to a biological problem for the first time. This stage is usually the most readily fundable by typical biology funding agencies, as long as the focus remains on the biological question, although reviewers vary in what they consider a trivial adaptation versus new methods development. In addition to requiring biological expertise, this work requires researchers who are aware of the latest computational advancements and can implement methods. In past decades, this often meant interpreting a narrative and equation-laden description of the algorithm and implementing it from scratch in a particular language – because, culturally, computer scientists have generally not prioritized providing executable code with their papers (although this custom is beginning to change substantially, especially in machine learning). More recently, implementing a method has often meant configuring machine-learning architectures and learning strategies. Often methods work right 'off the shelf', but sometimes adaptation of the core methodology is needed. For example, most machine-learning methods designed for natural images take in three-channel (red, green and blue) images, with substantial overlap of the signal in the channels. By contrast, biological images are frequently greyscale (for example, electron microscopy) or multichannel (often with four or five channels, but sometimes with dozens), with a dark background and very sparse light regions that have little signal in common with other channels. Adaptation can be quite time consuming, and it requires expertise that is often not found in biological research labs. Although this sort of research is rarely itself considered for publication in prestigious journals, it can drive ground-breaking biological discoveries.

Usable tools are the 'last mile' bridge between what computer science makes possible and what biologists are able to put to widespread use in their research. 'Last mile' problems exist in various industries: for example, public transit can inexpensively transport many people between two stations, but if people do not have a reasonable means for getting from stations to their final destination (the last mile), the transit system remains unused. Likewise, brilliant and useful computational methods may exist, but these potentially dramatic advances go to waste if they reside only in equation-dense papers, or in piles of poorly explained, messy code that most researchers do not know how to install, run or configure properly. Rarely are usable biological software bridges built, owing to the lack of dedicated resources and rewards for this stage; this work requires software engineering expertise, which is expensive and underfunded by funding agencies' typical mechanisms[7,8], because usability and maintenance work does not itself constitute new research.

Pioneering exceptions exist, such as the Chan Zuckerberg Initiative's 'Essential open source software for science' program; the 'Research software 2023' program of the Deutsche Forschungsgemeinschaft (German Research Foundation); Schmidt Futures' Virtual Institute of Scientific Software[9]; the UK's Software Sustainability Institute (run as an unusual consultancy model)[10]; the US National Science Foundation's 'Pathways to enable open-source ecosystems' program (among others); and various programs within the US National Institutes of Health (NIH), such as 'Administrative supplements to support enhancement of software tools for open science' and 'Biomedical technology optimization and dissemination centers'. Although they fall short of the scale of critical need – a European Union study found a major shortfall, especially for long-term maintenance – these funding sources have kept many crucial scientific software projects healthy enough for use. A further cultural

issue is that, often, pursuing the funding needed to create usable software is not a priority for researchers because the academic reward system tends to reward new contributions over useful ones. Some prestigious journals have begun special article types for software projects, mitigating this hurdle. Dedicating time to creating educational materials and training users is even less well rewarded but is crucial for the adoption of software[11,12].

Yet we contend that creating usable tools is a crucial step in the scientific process, with exceptional return on investment. ImageJ was started as a 'pet project' in 1987 by Wayne Rasband at the NIH and was maintained mostly by him for decades. It has been cited in more than 100,000 papers and did not receive an external NIH grant until 2009. I (A.E.C.) spent my postdoctoral fellowship and a substantial portion of my lab startup funds on CellProfiler's early development. Some felt it would be a fatal error to focus on relatively unpublishable user-friendly software in my early career. Although it may have made both funding my lab and publishing more difficult, its citation in more than 16,000 papers and its widespread use in academia and the pharmaceutical industry attests to its value to researchers. Large language models have been available in their modern form since the early 2010s, yet the frenzy about their power began only in late 2022, in large part due to more data and computing power, but in equal part because OpenAI introduced ChatGPT, making the technology accessible to the masses.

Clearly, computer science has transformed most fields of biology by making research faster, cheaper and more information rich. However, a need remains for dedicated resources and a shift in the academic reward system to create user-friendly tools that bridge the gap between computer science and biology[8,12]. The benefits of doing so are tremendous; for example, in our vision of a smart microscope of the future[13], artificial intelligence could in theory power plain-language-guided image acquisition, automatic image analysis based on extensive prior training from biologist experts, and plain-language-guided image analysis for custom analyses. Similarly, in medicine, generalist medical artificial intelligence models could power grounded radiology reports, assist with surgical procedures, provide bedside decision support and interactive note-taking, and power chatbots for patients[14]. Similar translation could bring the latest advances in artificial intelligence to all domains of biology and biomedicine, from bioinformatics analyses to literature summarization. Bridging the last mile will maximize the impact of the dizzying advancements in computer science, and drive biological discovery.

References

1. Min Jou W, Haegeman G, Ysebaert M & Fiers W Nature 237, 82–88 (1972). [PubMed: 4555447]
2. Lander ES et al. Nature 409, 860–921 (2001). [PubMed: 11237011]
3. Venter JC et al. Science 292, 1838 (2001).
4. 1000 Genomes Project Consortium. Nature 526, 68–74 (2015). [PubMed: 26432245]
5. Jumper J et al. Nature 596, 583–589 (2021). [PubMed: 34265844]
6. Ronneberger O, Fischer P & Brox T In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (eds. Navab N et al.) vol 9351, 10.1007/978-3-319-24574-4_28 (2015).
7. Anzt H et al. F1000 Res. 9, 295 (2020).
8. Deschamps J, Nogare DD & Jug F Front. Bioinform 10.3389/fbinf.2023.1255159 (2023).

9. Matthews D Nature 607, 410–411 (2022). [PubMed: 35831588]
10. Crouch S et al. Comput. Sci. Eng 15, 74–80 (2013).
11. Carpenter AE, Kamentsky L & Eliceiri KW Nat. Methods 9, 666–670 (2012). [PubMed: 22743771]
12. Soltwedel JR & Haase R J. Microscop 10.1111/jmi.13192 (2023).
13. Carpenter AE, Cimini BA & Eliceiri KW Nat. Methods 20, 962–964 (2023). [PubMed: 37434001]
14. Moor M et al. Nature 616, 259–265 (2023). [PubMed: 37045921]

Fig. 1 |.

Stages of technology-driven acceleration in biological research that involves computational methods.