ACS Synthetic Biology
Cite This: ACS Synth. Biol. 2019, 8, 1474−1477
Viewpoint
pubs.acs.org/synthbio

Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation

Pablo Carbonell,[†] Tijana Radivojevic,[‡,∥] and Héctor García Martín*[,‡,§,∥,⊥]

[†]Manchester Synthetic Biology Research Centre for Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology and School of Chemistry, University of Manchester, Manchester M1 7DN, United Kingdom

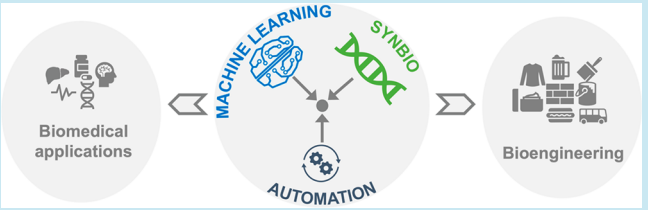
[‡]DOE Agile BioFoundry, Emeryville, California 94608, United States

[§]DOE Joint BioEnergy Institute, Emeryville, California 94608, United States

[∥]Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States

[⊥]BCAM, Basque Center for Applied Mathematics, 48009 Bilbao, Spain

ABSTRACT: Our inability to predict the behavior of biological systems severely hampers progress in bioengineering and biomedical applications. We cannot predict the effect of genotype changes on phenotype, nor extrapolate the large-scale behavior from small-scale experiments. Machine learning techniques recently reached a new level of maturity, and are capable of providing the needed predictive power without a detailed mechanistic understanding. However, they require large amounts of data to be trained. The amount and quality of data required can only be produced through a combination of synthetic biology and automation, so as to generate a large diversity of biological systems with high reproducibility. A sustained investment in the intersection of synthetic biology, machine learning, and automation will drive forward predictive biology, and produce improved machine learning algorithms.



A NEW BIOLOGY FOR A NEW CENTURY

Biology has changed radically in the past two decades, transitioning from a descriptive science into a design science. The discovery of DNA as the repository of genetic information, and of recombinant DNA as an effective way to modify it, has first led into the development of genetic engineering and later the field of synthetic biology. Synthetic biology[1] goes beyond the historical practice of a biological research based on describing and cataloguing (*e.g.*, Linnaean taxonomic classification or phylogenetic tree development), and aims to design biological systems to a given specification (*e.g.*, production of a given amount of a medical drug or targeted invasion of a specific type of cancer cell).

This transition into an industrialized synthetic biology is expected to affect most human activities, from improving human health, to producing renewable biofuels to combat climate change.[2] Some examples commercially available now include synthetic leather and spider silk, renewable biodiesel that propels the Rio de Janeiro public bus system, vegan burgers with meat taste, and sustainable skin-rejuvenating cosmetics.

In this effort, new tools enable us to bioengineer cells faster than ever: CRISPR-enabled genetic editing has revolutionized our ability to edit DNA *in vivo*, DNA synthesis productivity improves as fast as Moore's law, transcriptomics data volume has a doubling rate of 7 months, and high-throughput workflows for proteomics and metabolomics are emerging. Furthermore, the miniaturization and automation of these techniques through microfluidic chips[3] promise a future where data analysis rather than data production will be the bottleneck in biological research.

OBSTACLES TO AN EXPONENTIAL INCREASE IN SYNTHETIC BIOLOGY PRODUCTIVITY

However, despite new tools and exponentially increasing data volumes, synthetic biology cannot yet fulfill its true potential due to our inability to predict the behavior of biological systems. Arguably, the most pressing problems are our inability to predict the phenotype of biological systems when their DNA is altered, and the difficulty of using small scale experiments to predict the behavior at large scales.

In general, while we can make the DNA changes we intend on target cells, the end result on their behavior is usually unpredictable.[4] This limitation has led to a traditional bioengineering approach that involves randomizing experimental efforts hoping for an improved result, or using arduously gathered biological intuition. This approach is hardly scalable, and has resulted in long development times: for example, it took 150 person-years of effort for heterologous expression of the 16-enzyme artemisinin pathway, and 575 person-years of effort for DuPont's 1,3-propanediol.[5]

Special Issue: IWBDA 2018

Received: December 27, 2018
Published: July 19, 2019

DOI: 10.1021/acssynbio.8b00540
ACS Synth. Biol. 2019, 8, 1474−1477

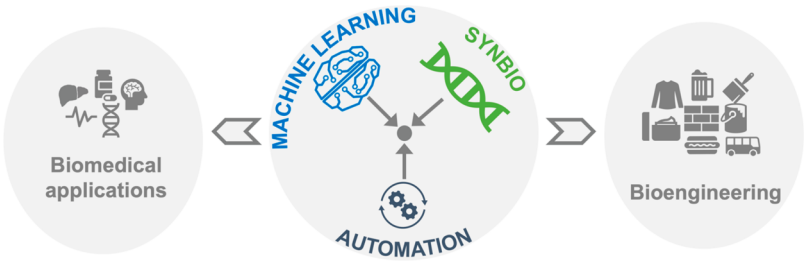


Figure 1. Synthetic biology, machine learning, and automation complement each other naturally. Combined, they can significantly increase our bioengineering capabilities and produce new biomedical applications.

Furthermore, we lack the ability to extrapolate large-scale behavior from small-scale experiments. In bioengineering, a key bottleneck is designing fermentation systems that reliably scale up lab results (1−100 mL) to commercial volumes ($100-10^6$ L). Failure to do so and meet production timelines resulted in the past in the inability to address high-volume production, economic losses, and significantly decreased investment in the field. Amyris, for example, had to announce major changes to its financing, strategy, and production targets after falling significantly short of their target of producing nine million liters of farnesene.[6] In biomedical applications, we cannot use information on cell culture experiments to reliably extrapolate the implications on human health. This shortcoming forces researchers to rely on proxy systems (animal models) such as mice, rats, pigs, monkeys, or rabbits. These animal models imperfectly represent human biology in biomedicine: the average rate of successful translation from animal models to clinical cancer trials is less than 8%. These failures significantly contribute to the billion dollar figures routinely cited for new drug development.[7]

While these two problems (predicting phenotype from DNA and scaled behavior) are perhaps more evident in the field of synthetic biology, they are shared with (and inherited from) the rest of biology. For example, it would be transformative to predict (1) plant phenotype from its genome, (2) soil microbial community impact on its environment and globally on Earth's climate from the study of pure cultures, or (3) mammalian metabolism from single cell studies. Any advance in these two problems will positively impact other subfields of biology. Further hurdles facing synthetic biology (*e.g.*, product extraction and downstream purification, supplement precursor cost, toxicity, long-term stability, reproducibility, cross-talk) are important, but less generally impactful if solved.

MACHINE LEARNING'S PREDICTIVE CAPABILITIES

Machine learning can provide predictive power without the need for detailed mechanistic understanding, by learning the underlying regularities in experimental data. Training data is used to statistically link a set of inputs to a set of outputs through models that are expressive enough to represent almost any relationship, without being encumbered by biased assumptions. In this vein, machine learning has been used to predict pathway dynamics, optimize pathways through translational control, diagnose skin cancer, detect tumors in breast tissues, and predict DNA and RNA protein-binding sequences.[8−10] Furthermore, machine learning can be used to design synthetic biology systems: it can be used to learn the relationship between phenotype and the genetic parts used in genetic circuits, thus allowing more stable circuits.

But machine learning algorithms are data-hungry: they require abundant data to be trained and be effective. The recent revolution in machine learning was enabled not by new algorithms, but rather by (1) growing computational power and (2) the availability of large training libraries. Image recognition in the field of artificial vision would have most likely not reached superhuman performance if it had to be trained on pictures captured on photographic film and physically mailed from photographers to artificial intelligence researchers. The availability of large image libraries enabled by automated digital image acquisition through charge-coupled device (CCD) cameras, and its dispersal through the Internet, have been key to its development.

MACHINE LEARNING NEEDS AUTOMATION TO BE TRULY EFFECTIVE

We cannot produce the quantity and quality of data needed for effective machine learning without using automation. The situation we face in biology is akin to using mailed paper pictures: most assays are low-throughput and manual, and most phenotypic data is produced and analyzed within the same lab. Although this is beginning to change, the rate is not fast enough to support machine learning approaches (except for the field of genomics). To make matters worse, historical data not always meet the requirements for machine learning to be effective (*e.g.*, lack of standardized data collection), so it is important that new data are collected with these needs in mind. Competitions such as the Critical Assessment of methods of protein Structure Prediction (CASP) provide a good example of how to promote community effort for this purpose.

Large-scale high-quality data is necessary but not sufficient: proper experimental design is fundamental to leverage machine learning. Opportunities in this area run in both directions: high-quality data generation for training machine learning algorithms necessitates experimental designs that carefully consider the different effects influencing the response; and machine learning can be used to choose the next set of experiments in order to improve experimental data quality and reduce the estimation errors. In this area, "robot scientists" (chemical experiment planners) have proven successful in synthetic chemistry, and are expected to play an important role in synthetic biology.

Hence, we need to invest in capabilities that couple machine learning algorithms with high-throughput, fast-turnaround, automated phenotyping approaches, to solve biological problems whose solution is of wide applicability (Figure 1). Possible approaches involve robotic liquid handler platforms, microfluidics, or cloud laboratories. Future challenges include acquiring data in real time, developing comprehensive

DOI: 10.1021/acssynbio.8b00540
ACS Synth. Biol. 2019, 8, 1474−1477

noninvasive assays, taking the human out of the loop, and developing workflows and data standards that ensure reproducibility.

While this approach is already being embraced in industry (*e.g.*, Amyris, Zymergen, Ginkgo, Genomatica), it would significantly benefit academic research. The availability of large amounts of high-quality data would enable computational biologists to produce robust theories without the need of running their own experimental facilities, and the theory produced by these data sets would allow experimentalists to better design experiments and tackle questions of general relevance. Furthermore, this division of labor would enable higher productivity and allow for addressing more ambitious biological questions. Indeed several academic biofoundries have recently appeared (*e.g.*, Edinburgh Genome Foundry, Illinois Biological Foundry for Advanced Biomanufacturing, Agile BioFoundry, Manchester SynBioChem, Tianjin University BioFoundry), which can provide the ideal environment for the integration of synthetic biology, machine learning, and automation, if properly directed and resourced (see Figure 2 for the role of machine learning and automation in the Design-Build-Test-Learn cycle).

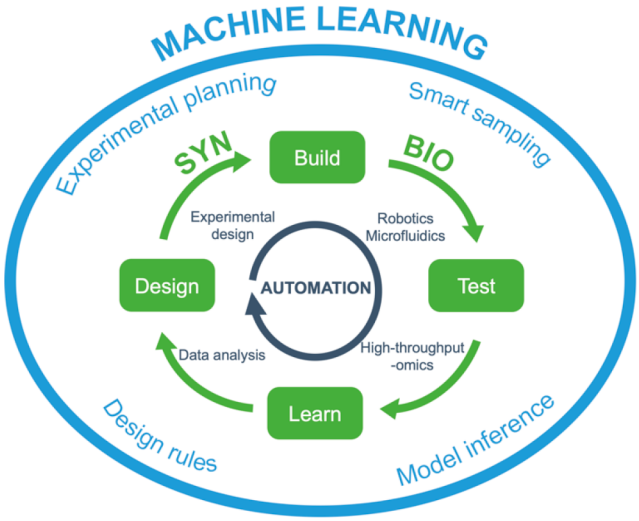


Figure 2. Machine learning and automation can be used to improve the basic synthetic biology Design-Build-Test-Learn (DBTL) cycle in different ways. Automation allows rapid growing and assembly of genetic designs through robotics and microfluidics platforms, high-throughput omics quantification, and experimental data analysis. Machine learning can drive each step in the cycle through generation of the experimental planning, smart selection of samples for quantification, model inference from experimental data, and design rules for the next iteration.

PREDICTIVE SYNTHETIC BIOLOGY WILL DRAMATICALLY IMPACT BIOLOGY AND INSPIRE COMPUTER SCIENCE

A significant opportunity lies in the integration of synthetic biology, machine learning, and automation, enabling disruptive changes in both biology and computer science. This integration can not only produce transformational synthetic biology applications for the production of biomaterials, biofuels and biomedical applications, but also enable a better mechanistic understanding. Unlike for other domains where machine learning is leveraged productively (*e.g.*, image recognition), for many of the current synthetic biology applications we have a significant (but not complete) knowledge of the underlying processes. Coupling the predictive ability of machine learning models with the possibilities afforded by new synthetic biology tools to easily modify the system components will allow us to probe and expand our mechanistic understanding. We expect this improved understanding to help us generate new types of machine learning algorithms: after all, machine learning staples such as genetic algorithms and artificial neural networks were inspired by biological analogies. This integration will require a tight multidisciplinary collaboration among biologists, mathematicians, engineers, chemists, physicists, and computer scientists in order to be successful.

AUTHOR INFORMATION

Corresponding Author

*E-mail: hgmartin@lbl.gov.

ORCID

Pablo Carbonell: 0000-0002-0993-5625

Héctor García Martín: 0000-0002-4556-9685

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

P.C. acknowledges the funding from the UK Biotechnology and Biological Sciences Research Council (BBSRC) and UK Engineering Physical Sciences Research Council (EPSRC) under grant BB/M017702/1, "Centre for synthetic biology of fine and speciality chemicals (SYNBIOCHEM)" and from BBSRC under grant BB/R506497/1 "Flexible Talent Mobility". T.R. and H.G.M. performed this work as part of the DOE Agile BioFoundry (http://agilebiofoundry.org), supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, and the DOE Joint BioEnergy Institute (http://www.jbei.org), supported by the Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). H.G.M. was also supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2017-0718.

REFERENCES

(1) Cameron, D. E., Bashor, C. J., and Collins, J. J. (2014) A brief history of synthetic biology. *Nat. Rev. Microbiol. 12*, 381−390.

(2) Committee on Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals, Board on Chemical Sciences and Technology, Board on Life Sciences, Division on Earth and Life Studies, National Research Council. (2015) *Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals*, National Academies Press, Washington, D.C., DOI: 10.17226/19001.

(3) Heinemann, J., Deng, K., Shih, S. C. C., Gao, J., Adams, P. D., Singh, A. K., et al. (2017) On-chip integration of droplet microfluidics and nanostructure-initiator mass spectrometry for enzyme screening. *Lab Chip 17*, 323−331.

DOI: 10.1021/acssynbio.8b00540
ACS Synth. Biol. 2019, 8, 1474−1477

(4) Gardner, T. S. (2013) Synthetic biology: from hype to impact. *Trends Biotechnol. 31*, 123−125.
(5) Hodgman, C. E., and Jewett, M. C. (2012) Cell-free synthetic biology: thinking outside the cell. *Metab. Eng. 14*, 261−269.
(6) *The Rise and Fall of the Company That Was Going To Have Us All Using Biofuels*, https://www.fastcompany.com/3000040/rise-and-fall-company-was-going-have-us-all-using-biofuels.
(7) Avorn, J. (2015) The $2.6 billion pill−methodologic and policy considerations. *N. Engl. J. Med. 372*, 1877−1879.
(8) Costello, Z., and Martin, H. G. (2018) A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *npj Syst. Biol. Appl. 4*, 19.
(9) Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542*, 115−118.
(10) Jervis, A. J., Carbonell, P., Vinaixa, M., Dunstan, M. S., Hollywood, K. A., and Robinson, C. J. (2019) Machine learning of designed translational control allows predictive pathway optimization in Escherichia coli. *ACS Synth. Biol. 8*, 127.

DOI: 10.1021/acssynbio.8b00540
ACS Synth. Biol. 2019, 8, 1474−1477