Easy all-in-one analysis, design, and interpretation of biological sequences with minimal coding

June 21 2023, by Lindsay Brownell

Credit: Harvard University

The amount of data generated by scientists today is massive, thanks to the falling costs of sequencing technology and the increasing amount of available computing power. But parsing through all that data to uncover useful information is like searching for a molecular needle in a haystack.

Machine learning (ML) and other artificial intelligence (AI) tools can dramatically speed up the process of data analysis, but most ML tools are difficult for non-ML experts to access and use. Recently, automated machine learning (AutoML) methods have been developed that can automate the design and deployment of ML tools, but they are often very

complex and require a facility with ML that few scientists outside of the AI field have.

A group of scientists at the Wyss Institute for Biologically Inspired Engineering at Harvard University and MIT has now filled that unmet need by building a new, comprehensive AutoML platform designed for biologists with little to no ML experience. Their platform, called BioAutoMATED, can use sequences of nucleic acids, peptides, or glycans as input data, and its performance is comparable to other AutoML platforms while requiring minimal user input. The platform is described in a new paper published in *Cell Systems* and is available to download from GitHub.

"Our tool is for folks who don't have the ability to build their own custom ML models, who find themselves asking questions like, "I have this cool data set, will ML even work for it? How do I get it into an ML model? The complexity of ML is what's stopping me from going further with this data set, so how do I overcome that?'", said co-first author Jackie Valeri, a graduate student in the lab of Wyss Core Faculty member Jim Collins, Ph.D. "We wanted to make it easy for biologists and experts in other domains to use the power of ML and AutoML to answer fundamental questions and help uncover biology that means something."

AutoML for all

Like many great ideas, the seed that would become BioAutoMATED was planted not in the lab, but over lunch. Valeri and co-first authors Luis Soenksen, Ph.D. and Katie Collins were eating together at one of the Wyss Institute's dining tables when they realized that despite the Institute's reputation as a world-class destination for biological research, only a handful of the top experts working there were capable of building and training ML models that could greatly benefit their work.

"We decided that we needed to do something about that, because we wanted the Wyss to be at the forefront of the AI biotech revolution, and we also wanted the development of these tools to be driven by biologists, for biologists," said Soenksen, a Postdoctoral Fellow at the Wyss Institute who is also a serial entrepreneur in the science and technology space. "Now, everyone agrees that AI is the future, but four years ago when we got this idea, it wasn't that obvious, particularly for biological research. So, it started as a tool that we wanted to build to serve ourselves and our Wyss colleagues, but now we know that it can serve much more."

While various AutoML systems have already been developed to simplify the process of generating ML models from datasets, they typically have drawbacks; among them, the fact that each AutoML tool is designed to look at only one type of model (e.g., neural networks) when searching for an optimal solution. This limits the resulting model to a narrow set of possibilities, when in reality, a different type of model altogether may be more optimal. Another issue is that most AutoML tools aren't designed specifically to take biological sequences as their input data. Some tools have been developed that use language models for analyzing biological sequences, but these lack automation features and are difficult to use.

To build a robust all-in-one AutoML for biology, the team modified three existing AutoML tools that each use a different approach for generating models: AutoKeras, which searches for optimal neural networks; DeepSwarm, which uses swarm-based algorithms to search for convolutional neural networks; and TPOT, which searches non-neural networks using a variety of methods including genetic programming and self-learning. BioAutoMATED then produces standardized output results for all three tools, so that the user can easily compare them and determine which type produces the most useful insights from their data.

The team built BioAutoMATED to be able to take as inputs DNA, RNA,

amino acid, and glycan (sugars molecules found on the surfaces of cells) sequences of any length, type, or biological function. BioAutoMATED automatically pre-processes the input data, then generates models that can predict biological functions from the sequence information alone.

The platform also has a number of features that help users determine whether they need to gather additional data to improve the quality of the output, learn which features of a sequence the models "paid attention" to most (and thus may be of more biological interest), and design new sequences for future experiments.

Nucleotides and peptides and glycans

To test-drive their new framework, the team first used it to explore how changing the sequence of a stretch of RNA called the ribosome binding site (RBS) affected the efficiency with which a ribosome could bind to the RNA and translate it into protein in E. coli bacteria. They fed their sequence data into BioAutoMATED, which identified a model generated by the DeepSwarm algorithm that could accurately predict translation efficiency.

This model performed as well as models created by a professional ML expert, but was generated in just 26.5 minutes and only required ten lines of input code from the user (other models can require more than 750). They also used BioAutoMATED to identify which areas of the sequence seemed to be the most important in determining translation efficiency, and to design new sequences that could be tested experimentally.

They then moved on to trials of feeding peptide and glycan sequence data into BioAutoMATED and using the results to answer specific questions about those sequences. The system generated highly accurate information about which amino acids in a peptide sequence are most important in determining an antibody's ability to bind to the drug

ranibizumab (Lucentis), and also classified different types of glycans into immunogenic and non-immunogenic groups based on their sequences. The team also used it to optimize the sequences of RNA-based toehold switches, informing the design of new toehold switches for experimental testing with minimal input coding from the user.

"Ultimately, we were able to show that BioAutoMATED helps people 1) recognize patterns in biological data, 2) ask better questions about that data, and 3) answer those questions quickly, all within a single framework—without having to become an ML expert themselves," said Katie Collins, who is currently a graduate student at the University of Cambridge and worked on the project while an undergraduate at MIT.

Any models predicted with the help of BioAutoMATED, as with any other ML tool, need to be experimentally validated in the lab whenever possible. But the team is hopeful that it could be further integrated into the ever-growing set of AutoML tools, one day extending its function beyond biological sequences to any sequence-like object, such as fingerprints.

"Machine learning and artificial intelligence tools have been around for a while now, but it's only with the recent development of user-friendly interfaces that they've exploded in popularity, as in the case of ChatGPT," said Jim Collins, who is also the Termeer Professor of Medical Engineering & Science at MIT. "We hope that BioAutoMATED can enable the next generation of biologists to faster and more easily discover the underpinnings of life."

"Enabling non-experts to use these platforms is critical for being able to harness ML techniques' full potential to solve long-standing problems in biology, and beyond. This advance by the Collins team is a major step forward for making AI a key collaborator for biologists and bioengineers," said Wyss Founding Director Don Ingber, M.D., Ph.D.,

who is also the also the Judah Folkman Professor of Vascular Biology at Harvard Medical School and Boston Children's Hospital, and the Hansjörg Wyss Professor of Bioinspired Engineering at the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS).

More information: BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences, *Cell Systems* (2023). dx.doi.org/10.1016/j.cels.2023.05.007

Provided by Harvard University