

Using the Artificial to Understand the Real:

A look at how Artificial
Intelligence is giving back
to the field of Biology

By Michael Mahan

Intro

Artificial Intelligence(AI) is based around the idea of replicating or even improving upon biological intelligence. So it comes to mind to question, what leaps and bounds has AI been able to return to biology, after many years of being influenced by it?

The answer to this question is returned in various fields, from drug development, to ecology, and even has helped cancer prediction/detection. In fact, it's so prevalent within modern biological science, that some are starting to voice concerns about whether it's being properly used by researchers, and properly scrutinized by reviewers. First though, let's see what's being done to mitigate human extinction from ecological change.

Ecology

Climate change is something that affects the ecological sphere, so even here we find motivation for AI in food production. As temperatures grow on global scale, we will going to see a lot of physical displacement of myriad peoples, as well as a large-scale reduction of farmable land. These two symptoms will be almost sequential: the lack of farmable land will undoubtedly cause this displacement, although rising sea levels will also certainly have a role to play. Regardless, we seem to have passed many estimated tipping points, and so the role of ecological science now has pivoted ever more towards making sure that food scarcity is avoided.

In 2015, a movement began through the UN, to develop plant life for human consumption, with the future development of climate change in mind. These plants are to be optimized towards growing in harsher environments, as well as to use minimal water in their production, and generate optimal nutrients. AI has been chosen as the technological method for optimizing several different major levels of plant production, in the hopes of meeting these goals, which have been categorized as "ideotypes"(Streich).

Within proteomics, we have analysis of protein production. We have the ability to 3D model proteins and their movements with high resolution, thanks to HPC(high performance computing). We also have the ability to protein-protein interactions which is useful for running analysis on protein docking(Streich). Protein docking concerns itself with simulating how proteins interlock to form protein complexes. Protein complexes are "a cornerstone of many (if not most) biological processes and together they form various types of molecular machinery that perform a vast array of biological functions." (Wikipedia, Protein Complex). The application of AI here is towards the testing of these complexes, with foreign molecules, to find phenotypic qualities.(Streich) Phenotypic qualities could relate to both naturally occurring and artificial introduced chemicals; how the plant processes concentrations of carcinogens in the environment, whether or not there are negative interactions with certain pesticides. So processing of these interactions through ML's predictive ability, saves both time and effort that might have been wasted on entirely manual testing.

Phenomics is the analysis of visibly detectable traits within an organism. Image processing can easily differentiate 100s of different phenotypic expressions in plants. One example of this being utilized: machine learning was leveraged to map relationship between phenotypic traits and seed yield within soybeans, leading to the ability to process images of soybean plants without needing to wait for the fruition of seeds (Parmley). It is suggested that in the future, it would not be far flung to imagine a swarm of camera mounted drones surveying fields mid-season, and using image processing to predict end-of-season yields, which would allow for adjustments towards optimal growth, far ahead of time.

One of the notable current shortfalls of image processing for plants, is that smaller scale recognition is underutilized. In most cases, the entire plant is analyzed, as opposed to separate components such as the leaves or stalk. There is some work being done towards this issue, and it is predicted that this "will under-pin efforts to develop food and bioenergy crop ideotypes with improved climate resilience"(Streich).

So, will we as a species, have hope for reliable food sources in the not-so-far-off future? Almost certainly, though there seems to be no one looking out for flavor optimization. It

unfortunately might be a quite gruel-heavy end result that awaits us. Salt production won't be in any jeopardy though, so at least there will be *something* to add some smatter of flavor.

Pharmacology

Drugs are a very common facet of modern life. Vaccines help with prevention(when we actually use them), and antibiotics and antivirals rid us of what's already within. A myriad of other classifiers can be used to describe the pills and solutions some of us imbibe daily to just maintain our health, both mental and physical. Modern medicine is the forefront of what has extended life expectancy for our entire species. And modern medicinal development is always trying to improve and expand. It just so happens, AI can help with that.

In drug development, there are few candidates that make it through clinical trials, an approximately 90% rate of failure. While there are many commercial options for sending candidates through gauntlets that are fairly efficient, turning to AI seems to be a better option for resource management: it takes even less time, and tends to be more accurate than the standard quick methods.

Additionally, drug development produces large amounts of numerical data. This is a favorable mark towards using ML methods for the prediction of properties, and physiological effects/effectiveness. For some perspective, of the many data resources available to researchers, there's PubChem, which acts as "public repository for chemical structures and their biological properties"(Zhu), and houses over 96 million compound profiles. PubChem has been integral in many studies, and often hand in hand with some form of ML.

A more specific case, DNNs(deep neural networks) are in usage now, replacing an older statistical method called QASR(quantitative structure-activity relationship), due to DNNs's better performance in predicting many facets of drug/body interaction. In particular, a toxicity model called DeepTox, which is a DNN based model, participated and came first in a competition held by the NIH(National Institutes of Health)(Zhu). This competition involved quite a few other models, some based on other ML types, others on more traditional statistical methods. There is a rising popularity of DNNs with regards to pharmacology, generally in prediction of biological effects on targets, although also in modeling, with limited success.

With regards to drug *modeling*, we see less of DNNs, and more of CNNs(convolutional neural networks). CNNs seem to lend themselves more towards image recognition, and are often used for compound identification on a molecular level. CNNs have also been used to mine biomedical studies for drug interaction data, highly successfully(Zhao). The primary assistance AI is providing within drug development seems to be more efficiency oriented, rather than enabling drug researchers to do things they previously couldn't.

Overall, it engenders confidence in scientific progress, to look and find what public data resources are available, and to see that they're well used. It's entirely likely that the current pandemic will find its solution in a number of different places, but unlikely that any of those places are untouched by AI.

Cancer

Within the area of personal medicine, nothing is so common yet serious, as cancer. Symptoms are not always consistent, and can wildly differ between specific forms, especially in earlier stages. By the time a detection is made, sometimes it is already far too late to remedy. The cultural psychological effect is powerful: many seek to romanticize or even symbolize cancer through plays, film, music, and even video games. So it is little wonder that this fundamental symbol of human frailty, the result of the body failing to excise its own rogue elements, is the focus of many researchers. It is also consequent that trying to use AI to predict cancer is hugely popular; the predominant paper for this section has over 1,100 citations.

Researchers have, for a few decades, sought to detect the earliest stages of cancer. As they have committed themselves to this endeavor, a large corpus of data has been developed. The idea now is to use ML in conjunction with this body of data, in order to predict outcomes of individual cancers after their discovery within a patient, as well as to create a better grounds

for determining if a patient has cancer in the first place. The predominant ML techniques employed have been: ANNs(Artificial Neural Networks), DTs(Decision Trees), SVNs(Support Vector Machines), and BNs(Bayesian Networks)(Kourou). ANNs are cited as being both a "Gold standard" and unfortunately also a "black-box". DTs are considered good for being easier to interpret. SVMs have been used to determine a classification of tumors being benign or malignant, based on age of patient and size of tumor. SVMs are also considered easier to catch misclassified data with. BNs are less predictors, more probability generators, but are seen as useful for that case. All four of these methods have been widely used within literature concerned with intersecting ML and various aspects of cancer.

Cancer prognosis prediction is complex, and describes a series of variables: Patient susceptibility, the chance of cancer returning, and how likely a patient is to survive. In one study of patient susceptibility, an ANN was used with a data set of 62,219 patient's mammograms, and associated demographics and various other datum, which resulted in a very high accuracy of 96.5% when subjected to 10-fold cross-validation(Ayer). This is a very high rate of accuracy, and might be overfitting, but the sample size is also notably large. With regards to cancer recurrence, a variety of classifiers were used and subsequently compared in another study. After each method, a 10-fold cross-validation check was used. The eventual conclusion involved comparing two variations of BN usage, both by itself and in combination with a CFS algorithm, in order to generate a consensus from the results. The BN w/ CFS providing a 91.7% accuracy in predicting a cancer regression(Exarchos). Again a high accuracy, however this time the sample size is much smaller, at 86 patients, the likelihood of misclassifying or some form of bias is much higher here.

Lastly, we have an example of patient survival prediction. This is arguably the most pertinent aspect of cancer prognosis, as it's the final result. Reliability on this front is a double edged sword: once we achieve high accuracy people will either have very little to hope for, or come away feeling secure in their own mortality. A study of non-small cell lung cancer patient data managed a fairly reasonable 83% accuracy rate with an ANN(Chen). At a sample size of 440, this is a fairly reasonable rate. The unfortunate thing here is that the study was cancer specific, likely too specific for applicability to other cancers. This is not likely to be atypical, the fact is simply that different cancers operate differently is likely to remain a larger roadblock within the science.

The hope is that in the future, not only will AI continue to determine prognosis and diagnosis, but that it will also be capable of solidly differentiating benign and malignant tumors, and better use gene signatures for predicting outcomes. In terms of lack, the field seems to want for larger data sets, and since that's fairly important in ML related algorithms, it may become a focal point in the future. That said, it's clear from the sheer number of studies done within the last few years, that strides are being made and with that more lives can be saved.

Post-translational Modifications

In an effort to find an intersection between pharmacology, cancer, and biological processes, we find ourselves staring at something rather small. Post-translational modification(PTMs) are a small-scale point in the protein synthesis process. Just after initial creation, proteins may be modified by natural processes, which generates a "mature protein product"(Wikipedia, Post Translational Modification). Now this particular subject might be a bit specific, or rather *zoomed in*, but it's highly integral in disease-oriented drug research, with such targets as Alzheimer's, Parkinson's and various cancers.

Classically, researchers have not necessarily trusted generated results for compounds, at least not results that did not also have some form of experimental confirmation. The problem with this is that the methods for quality compound analysis are highly expensive, or not even possible for certain proteins/compounds. Over time however, this reluctance has eroded. Bioinformatics and the research tools developed within the field, have become integral with simulating models, and in specificity, a large number of these tools in recent years, have come with adherence to a set of procedures, known as Chou's 5-steps rule. Without getting too far

into specifics, Chou's rules are a logical sequence that crystalize a process for creation of a predictive algorithm from vetted data and then subsequently making the tool/algorithm produced from the data publicly available through web services(Chou).

AI-based tools can easily align with this process. The specific ones made with Chou's rules in mind are consequently transparent, easy to repeat results with, and generally highly convenient for researchers. It's thanks to this that many AI based tools for PTM are accessible, and benefitting the research community. This no doubt will lead to quicker results and further cooperation from within the research community, towards resolving diseases that center themselves around PTMs.

Research Concerns

While AI is seeing tremendous adoption within biology, as well as many other fields. There are however, concerns about the fidelity of the models, and researchers' choices when creating them. In particular, there is concern from a Dr. Jones at UC London that Machine Learning techniques are being used inappropriately. To wit, "either they do not offer any improvement in comparison with existing methods or their experimental design is flawed; often both."(Jones), and so a point of order is marking out where ML is and isn't appropriate.

Within his article, Jones highlights what he considers the base requirements: a large amount of data, high dimensionality(that is, many, many variables), and data that is highly structured. There's definitely an issue using some types of biological data, simply because they're more categorical than numerical, and yet researchers make submissions, using ML, with this sort of data. It's almost a sort of 'cargo cult programming' issue. Because a person studied in both biological sciences and versed in programming is uncommon, the result in the biological field is often experienced biologists using the highly abstracted ML tools that are available, and getting close enough to something correct that they can go with it. This is not a jab, but rather highlighting a common oversight, that knowing and detecting the pitfalls of ML usage are not immediately apparent.

This leads to an interesting, somewhat co-morbid issue: reviewers may see that a study used ML, and so make certain assumptions about it being particularly well done, simply because it's the 'cool new thing'. Jones argues this should not be the case, and in fact there should be more to the critiquing of papers that use ML. Looking out for things such as population bias, which is easier to determine, but also data leakage. Data leakage is more common within biological samples, because data that seems independent, can actually be counterintuitively dependent. Jones cites a particular case: protein structure prediction. In order to protect against proteins that may be evolutionarily dependent, it not uncommon that a threshold of <25% shared sequence identity is a cutoff for determining independence. However, "there are many cases of related proteins with zero sequence identity." (Jones), and so even here we find entanglement that would be hard to manage, and is not ideal for an ML form of analysis.

Overall, there's a push for checklists, or other methods of ensuring that both researchers and reviewers may introduce better integrity to their ML-based work within biological science.

Closing

In closing, AI, and more specifically the different flavors of ML, is absolutely accelerating the field of biology. From speeding up and enhancing pharmacological predictions and testing, to making sure that food supplies in the future will maintain our species. Improving cancer diagnosis and prognosis, as well as survivability. Allowing the entire research community to benefit towards solving the questions of Alzheimer's and other diseases that have plagued human beings for so long. Sometimes, there needs be a bit of extra scrutiny applied, but overall, biology is accomplishing wonderful things, all thanks in part to how far we've come with AI.

Bibliography(in order of appearance)

1. Streich, Jared, et al. "Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals?." *Current opinion in biotechnology* 61 (2020): 217-225.
2. Wikipedia contributors. "Protein complex." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 19 Sep. 2020. Web. 26 Sep. 2020.
3. Parmley, Kyle, et al. "Development of optimized phenomic predictors for efficient plant breeding decisions using phenomic-assisted selection in soybean." *Plant Phenomics* 2019 (2019): 5809404.
4. Zhu, Hao. "Big data and artificial intelligence modeling for drug discovery." *Annual Review of Pharmacology and Toxicology* 60 (2020): 573-589.
5. Zhao ZH, Yang ZH, Luo L, Lin HF,Wang J. 2016. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics* 32:3444–53
6. Fu, Xiao. "Application of Artificial Intelligence Technology in Medical Cell Biology." *2019 International Conference on Robots & Intelligent System (ICRIS)*. IEEE, 2019.
7. Liu, Yufei, et al. "Wasserstein gan-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology." *Engineering* 5.1 (2019): 156-163.
8. Kourou, Konstantina, et al. "Machine learning applications in cancer prognosis and prediction." *Computational and structural biotechnology journal* 13 (2015): 8-17.
9. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited. *Cancer* 2010;116:3310–21.
10. Exarchos KP, Goletsis Y, Fotiadis DI.Multiparametric decision support system for the prediction of oral cancer reoccurrence. *IEEE Trans Inf Technol Biomed* 2012;16:1127–34.
11. Chen Y-C, Ke W-C, Chiu H-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput Biol Med* 2014;48:1–7.
12. Wikipedia contributors. "Post-translational modification." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 17 Jun. 2020. Web. 26 Sep. 2020.
13. Chou, K. C. "Artificial intelligence (AI) tools constructed via the 5-steps rule for predicting post-translational modifications." *Trends Artif Intell* 3 (2019): 60-74.
14. Jones, David T. "Setting the standards for machine learning in biology." *Nature Reviews Molecular Cell Biology* 20.11 (2019): 659-660.