3 Ways Big Data and Machine Learning Revolutionize Drug Discovery

Nov. 16, 2016 by Andrii Buvailo

The Internet media is trending now with numerous mentions of “big data”, “machine learning” and “artificial intelligence” all together destined to revolutionize pharmaceutical and biotech industries and the way drugs are discovered. These new technologies are believed to make drug discovery cheaper, faster, and more productive.

But how is “magic” supposed to happen, after all?

First, let’s review briefly some of the basic concepts in the heart of new technologies.

Big Data: Volume, Velocity, and Variety

The term “big data” by itself is more of a marketing nature. It describes an abstract concept of having large volumes of data obtained from various channels in multiple formats, which needs to be arranged in such a way, that it can be possible to quickly access, search, update, and analyze it to output useful information.

For example, a big data analytics project may attempt to forecast weather by correlating simultaneous data streams from myriads of different sensors, satellites and observation stations. All this information comes in different data types, including structured data, such as SQL database stores (tables of data with columns and rows); unstructured data, such as document files (satellite photos, for example); or streaming data from sensors.

A crucial point of the big data concept is, also, a speed at which the information must be processed and analyzed.

The computing power required to quickly process huge volumes and varieties of data can not be achieved via traditional data management architectures using a single server or a server cluster. This can potentially demand hundreds or thousands of servers that can distribute the work and operate collaboratively. It is virtually impossible for most of the organizations to create and maintain anything of this kind.

So, although the concept of big data has been around for a long time -- a term “information explosion” was first introduced as early as in 1941, according to the Oxford English Dictionary -- a real data processing revolution happened much later, when a public cloud computing emerged as a primary tool for hosting big data analytics projects.

Nowadays, a public cloud provider can store petabytes (103 terabytes) of data and scale up thousands of servers for as long as it is needed to accomplish the big data project. This is available for a reasonable price and can be utilized by any organization in the world.

Furthermore, some public cloud providers offer not only “place” for storing data, but also big data capabilities, such as highly distributed Hadoop compute instances, data warehouses, databases and other related cloud services.

One example of big data services in a public cloud is Amazon Web Services Elastic MapReduce, which supports the processing of large data sets in a distributed computing environment.

Machine Learning

Machine Learning algorithm is a computer program that teaches computers how to adjust themselves so that a human do not need to explicitly describe how to perform the task to be achieved by the computer. The information that a Machine Learning algorithm needs in order to adjust its own program to solve a particular task is a set of known examples.

One of the revolutionary things about machine learning is that it can learn computers how to perform complex tasks, which are hard or even impossible for humans to describe and instruct. For example, if you want to make the upper photo in the figure below look like a Picasso painting, you can easily do so in seconds using machine learning. All you need to do is to provide a set of Picasso paintings to train the machine.

This is called supervised machine learning, it is when the program needs some example data to learn.

There are other learning techniques, which do not require a training dataset, for example, learning by “trial and error” -- unsupervised machine learning.

Since 2012, a specific Machine Learning technique called Deep Learning has been taking the AI world by storm. It deals with Artificial Neural Networks of different architectures and specific advanced algorithms for their training. A progress made within just three years since 2012 is larger than computer scientists

had done in the preceding twenty five years on several key problems, including Image Understanding, Signal Processing, Voice Understanding, and Text Understanding.

Exponential progress in big data processing and machine learning has led to a point when a combination of both technologies opened up huge practical potential for a variety of use cases, including data security, financial trading, marketing personalization, fraud detection, natural language processing, smart cars, healthcare, and drug discovery.

Big Data and Machine Learning In Drug Discovery

To understand how big data analysis and machine learning algorithms can improve drug discovery outputs, let's review three stages on the way to successful medicines, where new technologies fit in best.

1. Understanding biological systems and diseases

In most cases, a drug discovery program can only be initiated after scientists have come to understand a cause and a mechanism of action behind a particular disease, pathogens or medical condition.

Without exaggeration, biological systems are the most complex in the world and the only way to understand them is to follow a comprehensive approach, looking into multiple organizational "layers", starting from genes and all the way to proteins, metabolites and even external factors influencing inner "mechanics".

In 1990, a group of scientists began the process of decoding the human genome. It took 13 years and was worth $2.7 billion to have the project finished. Often called the Book of Life, deciphering the genome would not have been possible without massive amounts of compute power and custom software.

The genome is sort of "instruction" for the organism saying which proteins and other molecules should be produced, when and why. Having a complete knowledge of the genome opens doors to a much deeper understanding of our body, what can go wrong with it and under what circumstances.

However, looking at just genetic information is not enough, since genome is more like a paper map of the world: although it tells where cities and villages are located, it does not tell, who the inhabitants of those cities are, what they are doing and how they live. To better understand what is going on, scientists have to go beyond the genome's one-dimensional view into a multidimensional one, linking the genome with

large-scale data about the output of those genes at specific times, in specific places, in response to specific environmental pressures. This is what is called “multi-omic” analysis.

“Ome” here refers to different “layers” of biological system: Genome - all the genes in the body, DNA; transcriptome - a variety of RNAs and other molecules responsible for “reading” and “executing” genome information; proteome - all the proteins in the body; metabolome - all the small molecules; epigenome - the multitude of chemical changes to the DNA and factors, including environmental, which dictate such changes.

Such multidimensional approach is very promising for understanding mechanisms of diseases, especially, such complex ones as cancer and diabetes. They involve a tangled web of genes, the influence of lifestyle factors and environmental conditions. Whether you smoke or exercise daily, — that influences when those various genes are turned on and off.

Research on biology systems generates enormous amounts of data, which needs to be stored, processed and analyzed. The 3 billion chemical coding units that string together to form a person’s DNA, if entered into an Excel spreadsheet line-by-line, would produce 7,900 miles-long table. The human proteome contains more than 30,000 distinct proteins that have been identified so far. And the number of small molecules in the body, metabolites, exceeds 40,000. Mapping data, originated from various experiments, associations, combinations of factors and conditions, generates trillions of data points of information.

This is where Big Data analysis and Machine Learning algorithms start to shine, allowing to derive hidden data patterns, find dependencies and associations unknown before. For example, a recently reported automated protocol for large-scale modeling of gene expression data can produce models that are predictive of differential gene expression as a function of a compound structure. In contrast to the usual in silico design paradigm, where one interrogates a particular target-based response, the newly developed protocol opens doors for virtual screening and lead optimization for desired multitarget gene expression profiles.

Founded in 2015, a bioinformatics startup Deep Genomics developed new machine learning methods that can find patterns in massive datasets and infer computer models of how cells read the genome and generate biomolecules.

Another company, a Boston-based biopharma startup BergHealth, uses artificial intelligence-based analytics platform to engage the difference between healthy and disease environment in patient biology. According to the company’s CEO Niven Narain, the model they are using does not exist anywhere in the

world: “We’ve taken the genomics, looked at the metabolites and lipids, proteins, the clinical data, the drugs patients have used, the outcome they have, to really map this full narrative of patients”, he says.

Palo Alto-based bioinformatics startup NuMedii, Inc uses exclusive Big Data technology, originally developed at Stanford University, to analyze large amounts of scientific data together with proprietary biological network-based algorithms to discover drug-disease connections and biomarkers that are predictive of efficacy. The startup is active in the field of Integrative genomics, network-based methods, large-scale machine learning, and chemoinformatics.

A bioinformatics startup Insilico Medicine, Inc, recently formed Pharma.AI division developing deep learned transcriptomics-, proteomics-, blood biochemistry-based biomarkers of multiple diseases, predictors of alternative therapeutic uses of multiple drugs and analytical tools for high-throughput screening.

A very interesting startup Envisagenics uses cloud-based big data analytics to extract biologically relevant RNA isoforms from raw RNA-seq data. This startup’s software technology helps discover new drug targets and biomarkers through splice isoform quantification coupled with predictive analytics, prioritize disease-related genes, and provide a well-supported list of targets.

The important practical goal of the above research on biological systems is to be able to identify a protein or a pathway in the body, a “target”, playing a major role in a mechanism of a particular disease. Then, it would be possible to inhibit or otherwise modulate the target by chemical molecules to influence the course of the disease.

2. Finding the “right” drug molecules

Once a suitable biological target has been proposed by scientists, it is time to search for molecules which can selectively interact with the target, stimulating the desired effect -- a “hit” molecule.

A variety of screening paradigms exists to identify hit molecules. For example, a popular High-throughput screening (HTS) approach involves the screening of millions of chemical compounds directly against the drug target. In fact, it is sort of “trial and error” method to find a needle in the haystack. This screening paradigm involves the use of complex robotic automation, it is costly and the success rate is rather low. But what is good about it, though, is that it assumes no prior knowledge of the nature of the chemical compounds likely to have activity at the target protein. So, HTS appears to be an experimental source of ideas for further research, and it provides useful “negative” results to be taken into account.

Other approaches include fragments screening and a more specialized focused screening approach -- physiological screening. This is a tissue-based technique looking for a response more aligned with the final desired in vivo effect as opposed to targeting one specific drug target.

In the pursuit of cutting costs of the above complex laboratory screens and increasing their efficiency and predictability, computational scientists advanced computer-aided drug discovery (CADD) approaches using pharmacophores and molecular modeling to conduct so-called “virtual” screens of compound libraries. In this approach, millions of compounds can be in silico screened against a known 3D structure of a target protein (structure-based approach); if the structure is unknown, it is possible to identify drug candidates based on knowledge of other molecules which are known to have activity towards the target of interest.

CADD is another promising area where big data analytics and machine learning algorithms can become “superstars”.

A cheminformatics startup Numerate applies novel machine-learning algorithms, at cloud scale, to the problems of small-molecule drug design. Numerate has created a new innovative drug design platform that can rapidly deliver novel leads with no need for a crystal structure and with very limited SAR data. The approach the company follows consists of modeling the phenomena that are critical to the success of hits, leads, and drug candidates. Than, Numerate applies models thus derived to explore large chemical spaces to find novel therapeutics.

Cloud Pharmaceuticals, a biotech startup which is focused on the use of artificial intelligence and cloud computing to search virtual molecular space and design novel drugs. The startup’s technology can perform highly accurate binding affinity predictions, derive chemical property filters for drug-like properties, predict safety, and synthesizability of virtual compounds.

Atomwise, a health tech startup in the Y Combinator business incubator, uses Deep Learning Neural Networks to discover new medicines, achieving astounding results in hit discovery, binding affinity prediction, and toxicity detection. Recently, the company was able to go through 8.2 million compounds to find potential cures for multiple sclerosis in a matter of days. In the other project, Atomwise’s artificial technology was able to repurpose some of the existing drugs to suppress Ebola. These drugs were intended for unrelated illnesses and their potential to treat Ebola was previously unknown. The company collaborates with MERCK and other high-profile biopharma organizations.

Similarly, a recently founded biotech startup TwoXAR uses DUMA™ Drug Discovery platform, based on artificial intelligence, to evaluate large public and proprietary datasets to identify and rank high probability drug-disease matches. The obtained matches can be used to cross-validate clinical research, repurpose existing medicines, or identify novel drug candidates for further clinical testing.

3. Pre-clinical testing

One of the reasons why the pharmaceutical industry is experiencing a crisis and such a decline in research and development is because animal testing of new drug candidates is not very representative of what the human outcome will be. Drugs fail at later stages and it costs huge money for investors and wasted time for companies. What is more crucial, it costs lives for patients.

New artificial intelligence algorithms and big data approaches are now being applied to simulate the activity of many drugs on many tissues at once, like in a “virtual” human.

Concluding Remarks

Just several years have passed since famous tech entrepreneur Marc Andreessen penned his famous “Why Software Is Eating the World” essay. Today, a new statement proves itself true: “Software Eats Bio”.

New computational technologies and machine learning algorithms are revolutionizing biopharmaceutical industry and the way how drugs are discovered. A systemic understanding of biological processes and mechanisms of diseases opens doors to not only better drug molecules but also the whole new concept of personalized medicine, which takes into account individual variability in environment, lifestyle, and genes for each person. Big Data and Machine Learning are the technologies behind the future of Precision Medicine...