Biotechnology Journal
www.biotechnology-journal.com

# Biotechnology, Big Data and Artificial Intelligence

*Arlindo L. Oliveira*

**Developments in biotechnology are increasingly dependent on the extensive use of big data, generated by modern high-throughput instrumentation technologies, and stored in thousands of databases, public and private. Future developments in this area depend, critically, on the ability of biotechnology researchers to master the skills required to effectively integrate their own contributions with the large amounts of information available in these databases. This article offers a perspective of the relations that exist between the fields of big data and biotechnology, including the related technologies of artificial intelligence and machine learning and describes how data integration, data exploitation, and process optimization correspond to three essential steps in any future biotechnology project. The article also lists a number of application areas where the ability to use big data will become a key factor, including drug discovery, drug recycling, drug safety, functional and structural genomics, proteomics, pharmacogenetics, and pharmacogenomics, among others.**

## 1. Introduction

Rapid developments in biotechnology and in information technology have occurred in parallel, over the last half a century, at a rate unparalleled in any other field. Moore's law, the experimental observation that the number of transistors in an electronic chip doubles every two years,[1] is simultaneously a result and a cause of the exponential developments in information technology, which have few parallels in other areas outside the computer industry. One of the exceptions is exactly the biotechnology area, where sequencing and other high-throughput instrumentation technologies have been developing at an exponential rate that overshadows even Moore's law.[2] The data available and the new experimental technologies developed in the last decades make it easier and cheaper to perform experiments that would have taken years to undertake, just a few years ago. Many of these experiments will also generate large amounts of data, contributing to the data deluge that, at times, may overwhelm even the most data-savvy researchers.

The large amounts of data that are generated and stored, in the biotechnology area, create a range of new opportunities for researchers, as well as for companies that offer products and services in this area. However, it also creates challenges for researchers who are used to basing their work mainly on experimental data obtained locally, using low-throughput methods. Today, it is almost impossible to perform state-of-the-art research in biotechnology without using database and artificial intelligence technologies to process, explore, and exploit the vast troves of data available, public and private. In this article, I will offer a perspective of the methods that can be used to integrate data, to explore and exploit the information extracted from the many sources available, and the areas of biotechnology where these insights can become useful, or even essential.

Prof. A. L. Oliveira
INESC-ID, Instituto Superior Técnico
University of Lisbon
R. Alves Redol 9
1000-029, Lisboa, Portugal
E-mail: arlindo.oliveira@tecnico.ulisboa.pt

**DOI: 10.1002/biot.201800613**

## 2. The Rapid Growth of Biological Data

Empirical data suggests that nucleotide and proteomics data generation is growing at an exponential rate. Researchers at the European Bioinformatics Institute (EMBL-EBI) have monitored the growth of the different types of data stored in their servers and concluded that it is doubling roughly every year[3] (see **Figure 1**). This exponential growth raises significant challenges to the organizations that have the mission of storing and organizing such data but provides a bounty for biotechnology researchers who have the ability to explore and exploit the ever-increasing information stored in servers, worldwide. The (nonexhaustive) list of databases in the online molecular biology database collection kept by the *Nucleic Acids Research* journal includes more than 1700 operating databases.[4] These databases include information about nonvertebrate genomics (280 of them), protein sequence (214), human genes and diseases (176), molecule and protein structure (172), metabolic and signaling pathways (168), DNA nucleotide sequence (157), plants (131), vertebrate genomes (116), RNA sequence (104), microarray and gene expression (60), immunology (31), proteomics (28), organelles (20), cell biology (8), and other molecular biology information (92).

Many other databases, from related fields, such as medicine and chemistry, number in the thousands and are also relevant to biotechnology practitioners. Additionally, critically important information exists in electronic health records (EHRs) and other sources of medical data. Hospitals are especially keen on monitoring and assessing patient progress and their response to treatment plans. Companies such as Genentech and Predilytics, among many others, have developed extensive patient records, which store valuable information.

 

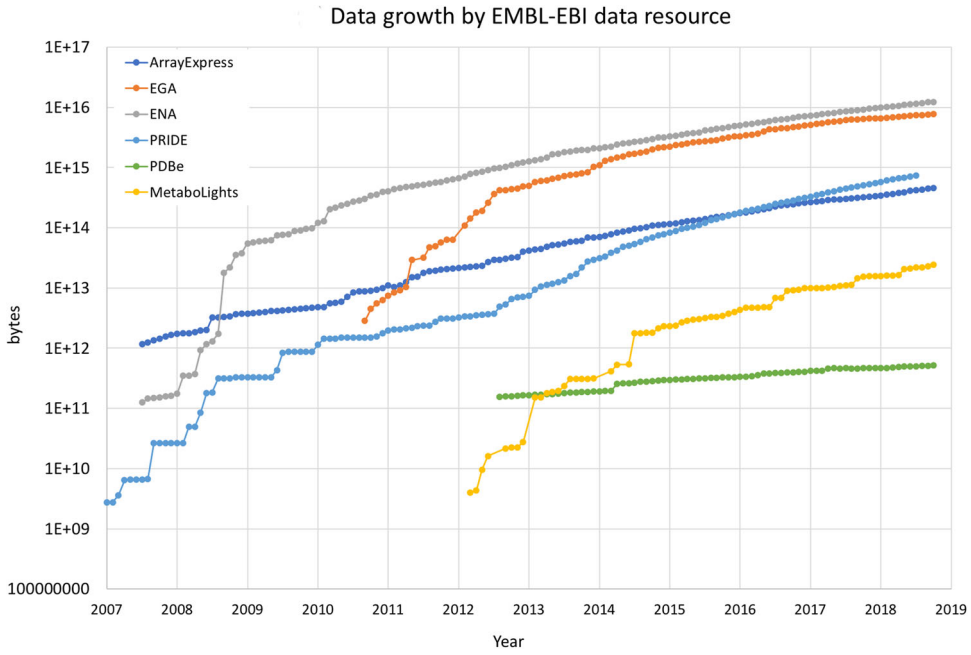


**Figure 1.** Growth of data stored by EMBL-EBI, for different types of data. Reprinted from Cook et al.,[3] with permission.

The prompt availability of all this information makes it necessary to use automated tools to integrate data from different sources, process and explore the relevant information in order to generate new knowledge and use this knowledge to develop and tune new processes, compounds, products, and services. In fact, it is almost unthinkable that any significant breakthroughs in biotechnology will be obtained in the future without resorting to the use of data stored online, in public and private databases.

Historically, the most significant results obtained in the biotechnological area have been obtained by laboratory experiments which, in many cases, used relatively limited amounts of data. The historic and well-known results in this area, ranging from Pasteur's discoveries about the fermentation process to Fleming's discovery of penicillin, were obtained by performing and analyzing in vitro experiments, which generated a pitiful amount of data when compared with today's large-scale experiments. This ability to work with small amounts of private data remained mostly true through the last decades of the 20th century. Until then, it was perfectly reasonable to assume that significant biotechnology research could be performed without resorting to the use of third-party data generated from large-scale experiments. However, with the sequencing of the human genome and the resulting outpour of data generated by sequencing, proteomics, and other high-throughput instrumentation technologies, it became rapidly unfeasible to make significant advances in this area without using big-data-based approaches.

Mining available data and, in some cases, publicly available data, has become one of the mainstays of biotechnology research. Yet, not every researcher in this area is familiar with the many tools that are now available to explore the vast amounts of knowledge dispersed through the thousands of databases that exist. Important new insights can be gathered by integrating, exploring, and exploiting information. In most cases, the complete process includes the integration of information from a number of different sources, the application of statistical and machine learning techniques to explore and extract knowledge from this integrated information, and the use of this knowledge to design, control, and improve bioprocesses.

## 3. Data Integration

Usually, the first step of data analysis involves the collection of information from a number of different sources and the integration of this information with locally generated data. Although it is quite possible to generate important new results by simply integrating and exploring existing public data, in most cases the incorporation of new data that is the result of the researcher's unique own background and experience makes the whole dataset more useful.

Data integration is a process that consists of retrieving, cleaning, and organizing data, usually obtained from a number of different sources. In most cases, a large fraction of the relevant data can be obtained from existing databases or from local databases that store relevant information, such as DNA, RNA, and protein sequences, gene expression, phenotypic information, protein structural data, protein-protein interaction, and protein expression. Additional information may come from signal transduction and metabolic pathway data, cell imaging data, and other sources. Medical information, obtained from the patient's EHRs and imaging data, can sometimes be a key factor in the success of projects in the pharmaceutical area. All these datasets need to be integrated and organized in order to become useful for a given objective.

Data integration approaches can be classified either as "eager" or "lazy."[5] In the eager approach, data is usually copied to a centralized data storage facility. In the lazy

approach, the data stays with the original sources and is integrated, on demand, whenever it is needed. Both approaches rely on the availability of machine-readable interfaces for the data, since downloading and organizing vast amounts of data by hand is challenging and, in most cases, prohibitive.

Many databases provide these machine-readable interfaces, known as web services, allowing users to perform automated requests for data, which can then be integrated, using either an eager or lazy approach. Web services can use a variety of protocols, including Asynchronous JavaScript And XML (AJAX), REpresentational State Transfer (REST), or Simple Object Access Protocol (SOAP). The exchanged information is transmitted in a machine-readable format, using one of the many available languages, such as the Extensible Markup Language (XML) or JavaScript Object Notation (JSON).

Genes, proteins, and all sorts of biological entities are referred to by researchers by a variety of names, leading to a lack of standardization that makes data integration harder. Since there are many ways to represent biological data, a number of standards have been proposed to define specific terms and structures for the representation of biological entities. Standards make it possible to integrate and conflate data from distinct sources. Despite its advantages, common and widely followed standards for biological data description and interchange are rare, and many different standards coexist, leading to significant difficulties. This situation is being addressed by a number of large-scale projects, including ELIXIR,[6] a pan-European project that unites Europe's leading life science organizations in the management and safeguard of the ever-increasing volume of data being generated by publicly funded research.

Once downloaded and integrated with the locally available information, data needs to be organized and structured. The technical options available to perform this step are numerous and reviewing them is outside the scope of this article. Popular options to store data are relational databases,[7] data warehouses based on star schemas,[8] and resource description framework (RDF) schemas.[9] Relational databases can model the underlying problem domain accurately but may become slow when answering some complex queries. Star- and snowflake-based data warehouses are built in order to answer quickly and efficiently a (more restricted) set of queries and are required when the amounts of data available are very large. RDF schemas, stored in databases called triplestores, can be used to answer semantic queries, and are, in some cases, a valid alternative to both relational databases and star-based data warehouses.

In some cases, machine-readable interfaces to publicly accessible databases are not available. In this case, researchers have to resort either to asking the database authors for database dumps or to use crawlers that can mimic the online behavior of human users and extract the data from websites, in a format (usually HTML) that can be later processed to extract the relevant data. Using crawlers, however, as well as database dumps, may raise significant intellectual property issues, which are usually taken care beforehand when web services make the data available, under prespecified conditions.

The main objective of the data integration steps is to organize the data in such a way as to make it easily usable, exploitable and, possibly, redistributable. In many cases, just organizing the data in a novel and easily exploitable way may represent a major contribution and create significant value, as is readily clear from even a cursory inspection of the databases listed in the successive editions of the *Nucleic Acids Research* database issues. However, the data is only useful if it can be effectively explored in order to obtain new knowledge. This means that it should be made either publicly available or explored by its owners in a deep and complete fashion.

## 4. Data Exploration

The objective of the data exploration phase is to extract new, hitherto undiscovered, knowledge that can be used to create new processes and products. Data exploration can be done in the old fashioned way, using human intuition and manual exploration of the data, but this approach is becoming rapidly unfeasible, with the fast growth of available data. Researchers must resort, therefore, to the use of a vast set of statistical and artificial intelligence techniques, all of which can be viewed as belonging to the wide area of machine learning. Data can be explored in a wide variety of ways using machine learning but two independent and complementary approaches may, in general, be used: supervised and unsupervised

In supervised machine learning, the objective is to infer a general rule from a set of labeled instances. In its simplest form, machine learning algorithms are provided with a table of instances, each of them labeled as belonging to a given class. The algorithm then infers the rules that make it possible to derive the label from the data. To provide one concrete example, imagine one has access to a database of gene expression levels in cells, where each cell is labeled as either cancerous or normal. This table of gene expression values together with the cell labels is called the training set since it is used to train the machine learning algorithms. Supervised machine learning methods can then be "trained" to infer the classification rules that classify a new, unseen, cell characterized by specific gene expressions values, as belonging to either the class normal or cancerous. These classification rules output the cell label when provided with the levels of gene expression of that one cell, which was not included in the original training set. The applications in biotechnology of supervised machine learning are almost endless, as witnessed by the steady stream of articles that apply it to biological problems. Expression data is not the only data that can be used in this particular problem, cell labeling. The available information, used to obtain the label of the cells, can be of many forms: expression data, DNA or protein sequence data, imaging data, or any other experimental records. Supervised machine learning methods are very flexible and can be used to label these cells, as long as there is available a set of previously labeled cells. This is just one particular application of supervised learning, used to illustrate the potential of the approach, but the reader should be able to extrapolate the key idea to other applications.

As machine learning technology evolves, most notably with the development of deep learning methods, high-dimensional data, such as images or videos,[10] can be processed automatically, at a rate that far exceeds the one that is possible to reach by using human beings to process the data. Many other applications of supervised learning exist and several will be referred later in this article. Many techniques can be used to perform supervised

machine learning, from regression and decision trees to support vector machines and deep neural networks.[11] Such approaches are usually called "classifiers," since they are trained to classify instances into one of a discrete set of classes. It is also possible to use these methods to address regression problems, cases where the labels are not discrete but continuous. Describing these techniques is outside the scope of this article, but all of them aim at building a model of the data that can be used to predict the label of previously unseen instances.

Given the current state of the technology, researchers do not need to reimplement these methods, as used to be the case just a few decades ago. Many publicly available packages exist and can be used directly to apply supervised machine learning methods to the user data. Among these, open-source packages like Keras,[12] a high-level interface to deep learning packages (Theano, Microsoft Cognitive Toolkit, and Tensorflow), Scikit-learn,[13] Torch,[14] and Weka,[15] among many others, are easy to use, competing directly with many commercial products from companies such as SAS, Microsoft, and IBM. Some of these packages require programming (using languages like R, MatLab, or Python) while others can be used by anyone with only minimum familiarity with data manipulation operations. Additionally, many companies, too numerous to list, develop and sell vertically integrated software packages for specific biotechnology applications. It is likely that we will assist, in the next decade, to the emergence of a small number of reference packages in the area, targeted at specific areas.

Unsupervised machine learning methods can also be used to explore the data but they are, in general, less powerful than supervised methods. Unsupervised machine learning methods do not require labels. Instead, they aggregate the instances provided in the training set, by similarity, in accordance with their properties. Clustering is a well-known unsupervised machine learning method, but there are several other approaches that perform a similar task. Clustering may lead to the discovery of regularities that may discover new knowledge, which was previously unknown. For instance, an unsupervised learning method, given a set of proteins may rediscover, by itself, that proteins can be partitioned into two different categories: hydrophilic and hydrophobic. Such a discovery may be the result of clustering the proteins, in accordance with the values of the properties used to characterize each one of them. In a similar way, such a group of proteins could be clustered in a different set of categories, possibly leading to the discovery of some new, hitherto unknown, property. However, given the undirected way unsupervised machine learning algorithms work, the discovery of new, relevant knowledge, by this type of algorithms, is less likely. For this reason, supervised machine learning algorithms are the most promising approach to use when looking for new knowledge.

## 5. Development and Optimization of Bioprocesses

Biotechnology is concerned not only with the creation of new knowledge about biological systems but also with the development of efficient bioprocesses that use this knowledge. In the previous section, I described how machine-learning techniques can be used to generate new knowledge, by integrating and exploiting data. In many cases, effective use of this knowledge in new products rests on the development of effective processes to synthesize them. Developing new processes has many phases and components, but several of them can also effectively benefit from the use of big data-based approaches.

The optimization of a bioprocess is a complicated task, as the efficiency of the process depends heavily on many factors, which are hard to optimize simultaneously. In fact, manufacturing costs are a major component of drug cost, and quality control techniques in the pharmaceutical industry still lag behind what is possible in other industries. The number of control variables involved in complex processes may number in the hundreds and finding the right combination of variables can be a daunting task. For example, what effects may small changes in the granulation, drying, milling, coating, or other manufacturing steps have on the quality and quantity of the final product? Traditionally, simultaneous optimization of all these degrees of freedom is not even attempted and bioprocesses are controlled with parameters that are known to lead to some reasonable process efficiency.

Understanding, predicting, and making more efficient bioprocesses is a tall order but has been made easier by the advent of new sensor technologies and big data approaches. Today, it is possible to instrument reactors and other devices in such a way as to obtain real-time information about the status of bioprocesses. Information from many sensors, including ambient sensors, cameras, probes, and sequencers, can be integrated in order to control and optimize, in real-time, processes that would otherwise run with only minimal adjustments. Analytics can be used to design, analyze, and control, in real-time, bioprocesses, given the data gathered from these sensors. Significant benefits may come in the form of faster development of new products, higher yields, shorter reactor cycle times, and reduced waste. However, having more data, by itself, does not mean better process control. There is a need to effectively integrate all the data coming from the instruments and to use it to optimize the process.

The basic idea is to develop models of the bioprocesses, which can be used to fine-tune the parameters that control these processes. Mathematical models for bioprocesses have been developed and extensively used to help control the reactors and production machinery. However, for complex processes, creating a good mathematical model can be difficult and, in many cases, undoable. Here, again, the application of machine learning techniques to data extracted from processes can be instrumental.

Supervised machine learning techniques, described in the previous section, can be used to infer the impact of a change in parameters, once trained with historical data about the process. In some cases, more complex machine learning techniques, such as recurrent neural networks, and long short-term memory (LTSM) networks[16] may be required to effectively model the effects of external variables on the behavior of bioprocesses, taking into account the process dynamics. Unlike physically realistic models, which try to model the physical variables of the process, machine learning models use a "knowledge-free" approach, where the physical reality of the process is ignored and only its input-output behavior is considered. Surprising as it may seem, these mechanisms

can in many cases model very accurately the response of complex systems, once trained in historical data obtained from the bioprocesses. The amount and quality of historical data are critical factors for the success of this approach. For this reason, many industries are now storing large amounts of data, for possible future use in process control. In fact, the whole concept of "Industry 4.0," a hot keyword these days, is to integrate data from many sources in order to optimize processes and achieve economic efficiency. Although the term has not been used extensively, it is reasonable to talk about "Biotechnology 4.0," the application of the very same ideas to bioprocesses.

## 6. Applications

The applications of big data in the biotechnology area are too numerous to even consider fully listing them here. However, it is useful to enumerate just a few potential areas of application, mostly to provide the reader with a few starting points to be further researched.

Experiment design is one area that can benefit from the usage of big data. The potential of machine learning to perform cell classification was already referred above, but, in general, all experiments should be designed having in mind the need to obtain, store, and process experimental data that can, in many cases, be analyzed automatically. Labor-intensive experiments can be automated, up to some degree, by carefully designing the experiments in such a way as to make the results of the experiment easy to monitor continuously, by automated computerized systems.

Drug discovery, performed by scanning databases for potential new drugs, is another promising area. New candidates for drugs can be identified by training a classifier on a dataset where functioning and nonfunctioning drugs have been identified. In a similar way, other properties of complex molecules can be determined by using classifiers trained on molecules with known properties. Machine learning techniques can be used to create virtual assays, identifying promising new drugs, which can later be tested in the laboratory. Identifying whether a new molecule affects a given pathway, or eliminates a given pathogen, can be done by using classifiers trained on other molecules, when the appropriate set of properties is used.

Mining EHRs is a process that can also provide useful information to be used in drug discovery, drug recycling, and drug safety research. Given the current focus on evidence-driven medicine, EHRs will become one of the most valuable resources hospitals and health institutions have to manage and explore, with the help of the biotechnology industry and academy. Both structured and nonstructured data in EHRs can be mined, the latter requiring the use of natural language processing technologies that are just now becoming of age. Future developments in pharmacogenomics will be strongly dependent on the ability of companies to integrate the information in EHRs with genetic information of the patients, which will become progressively more common.

In genome analysis, including Genome Wide Association Studies (GWAS), machine learning can be used not only to infer genotype-phenotype associations but also to identify the relations between genetic characteristics and the response to specific treatments. With the rapid decrease in sequencing and genotyping costs, more and more people are becoming interested in exploring their own genome and the consequences specific characteristics of this genome may have on their lifestyle. Companies like 23andme and Veritas Genetics, among many others, aim at collecting genetic data from millions of individuals, using it to create new knowledge, and fostering the development of new products. The 1000 Genomes project, launched in 2008, and the Million Genomes Project, launched in Europe in 2018, are just two of the many research projects aimed at collecting and making available large amounts of individual genome information. The availability of such large databases will make it possible to run in silico experiments that would have been unthinkable just a few decades ago.

## 7. Conclusions

Big data, artificial intelligence, and machine learning will become instrumental in all future biotechnology research. More and more researchers in biotechnology will have to become aware of the methods required to deal with large amounts of data and to include in their research teams people with the ability to integrate, organize, and explore this data. Researchers specialized in bioinformatics will become a key element in any biotechnology research team but, in many cases, the identification of opportunities will depend on the awareness about this topic of the biotechnology researchers themselves. This article aims at raising this awareness and, in this way, contributing to the development of this area of science.

## Acknowledgments

The author thanks the organizers of the 12th European Symposium of Biochemical Engineering Sciences, for the invitation to deliver an invited talk about the topics covered in this article, which led to the present article.

## Conflict of Interest

The author declares no conflict of interest.

## Keywords

artificial intelligence, big data, bioengineering, machine learning

Received: December 6, 2018
Published online: May 27, 2019

[1] G. E. Moore, in *Electronics Devices Meeting*, Washington, DC, December **1975**, pp. 11–13.
[2] E. Pettersson, J. Lundeberg, A. Ahmadian, *Genomics* **2009**, *93*, 105.
[3] C. E. Cook, M. T. Bergman, G. Cochrane, R. Apweiler, E. Birney, *Nucleic Acids Res.* **2015**, *46*, D21.
[4] D. J. Rigden, X. M. Fernández, *Nucleic Acids Res.* **2018**, *46*, D1.
[5] V. Lapatas, M. Stefanidakis, R. C. Jimenez, A. Via, M. V. Schneider, *J. Biol. Res.* **2015**, *22*, 9.
[6] L. C. Crosswell, J. M. Thornton, *Trends Biotechnol.* **2012**, *30*, 241.
[7] E. F. Codd, *Commun. ACM* **1970**, *13*, 377.

 

[8] R. Kimball, M. Ross, *The Data Warehouse Toolkit: the Complete Guide to Dimensional Modeling*, John Wiley & Sons, Indianapolis, IN **2011**.
[9] O. Lassila and R. R. Swick, W3C Recommendation 22 February 1999, **1998**, https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
[10] S. Webb, *Nature* **2018**, *554*, 555.
[11] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[12] F. Chollet, Keras, *GitHub Repository*. GitHub **2015**.
[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[14] R. Collobert, S. Bengio, J. Mariéthoz, *Idiap* **2002**.
[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10.
[16] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735.

 